UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot · your knowledge partner

THE MEASUREMENT INVARIANCE AND MEASUREMENT EQUIVALENCE OF THE SOURCES OF WORK STRESS INVENTORY (SWSI) ACROSS GENDER GROUPS IN SOUTH AFRICA.

by

Samantha Davis

Thesis presented in partial fulfilment of the requirements for the degree of Master of Commerce in the Faculty of Economic and Management Sciences at Stellenbosch University

Supervisors: Prof CC Theron & Dr G Görgens

December 2014

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signed:

Samantha Davis **Date:** 4 October 2013

ABSTRACT

The primary goal of an organisation, in a capitalistic system, is the maximisation of profit. The task of the human resource function in organisations is to affect the work performance of working man to the advantage of the organisation and in a manner that adds value to the organisation. The management of employee wellbeing/psychological health is one of the human resource interventions with which the human resource function pursues this objective. It is imperative for organisations to be aware of, and sensitive to, negative factors in the workplace, such as occupational stress, that influence employees' health and wellbeing and have a significant effect on job satisfaction and performance (Hamidi & Eivazi, 2010). Prevailing stress levels need to be monitored regularly if escalating stress levels are to be detected in time to prevent serious personal and organisational problems from developing. The Sources of Work Stress Inventory (SWSI) is an instrument developed in South Africa specifically for this purpose (De Bruin & Taylor, 2005). The inappropriate use of occupational stress assessments across genders can seriously jeopardize the extent to which occupational stress assessments, and the decisions based on them, achieve their intended objectives. In order to avoid making widespread generalisations and untested assumptions which will eventually do a disservice to the field of psychology, the absence of measurement bias (i.e. invariance and equivalence) should be demonstrated instead of simply assumed (Van de Vijver & Tanzer, 2004). Establishing the measurement invariance and equivalence of an instrument across groups should be a prerequisite to conducting substantive cross-group comparisons (Dunbar, Theron & Spangenberg, 2011). It is imperative to empirically ascertain whether the instruments that are used are free of cultural, language, gender, age and racial bias, not only because it is prohibited by the Employment Equity Act 55 of 1998, but also as it is in the interest of good workmanship. Bias is indicated as nuisance factors that threaten the validity of cross-group (cultural) comparisons (Van de Vijver & Leung, 1997). These nuisance factors could be due to construct bias, method bias and/or item bias. Due to the importance of the decisions made, it would seem essential that the information provided by test results apply equally across different reference groups. In this study the specific measurement invariance and equivalence sequence of tests set out by Dunbar et al. (2011) was used to answer a sequence of research questions that examine the extent to which the SWSI multi-group measurement model may be considered measurement invariant and equivalent or not, and to determine the source of variance if it existed (Vandenberg & Lance, 2000). Upon investigating the measurement model fit of the SWSI, the results indicated that support was found for the hypotheses that the measurement model fits the data of both gender samples independently. Furthermore, support was found for the configural and weak invariance model. However, due to not

meeting the requirements for metric equivalence, partial measurement invariance and equivalence was explored. The SWSI multi-group measurement model met the requirements of partial complete invariance and partial full equivalence, and the non-invariant items were identified in the process. The implications of the results are discussed, limitations are indicated and areas for further research are highlighted.

OPSOMMING

Die kerndoelwit van enige organisasie, veral in ‘n kapitalistiese stelsel, is om optimale wins te genereer. Die taak van die menslike hulpbronbestuurfunksie binne organisasies is om die werks-verrigting van die werkende mens te beïnvloed tot voordeel van die organisasie en terselfdetyd waarde tot die organisasie toe te voeg. Die bestuur van ‘n werknemer se welstand / sielkundige gesondheid is een van die menslike hulpbron-iintervensies waarmee die menslike hulpbronfunksie hierdie doelwit nastreef. Dit is uiters belangrik vir organisasies om bewus te wees van, asook sensitief te wees vir, negatiewe faktore soos werkstres, wat werknemers se gesondheid en welsyn beïnvloed en wat 'n beduidende invloed op werkstevredenheid en prestasie het (Hamidi & Eivazi, 2010). Heersende stresvlakke moet gereeld gemonitor word om tydig stygende stresvlakke te bespeur ten einde ernstige persoonlike en organisasieverwante probleme te verhoed. Die Bronne van die Werkstres-inventaris (BWSI) is in Suid-Afrika spesifiek vir hierdie doel ontwikkel (De Bruin & Taylor, 2005). Die ontoepaslike gebruik van werkstresmetings oor geslagte kan egter die mate waartoe beroepstresmetings en die besluite wat daarop gebaseer word hul oogmerke bereik ernstig benadeel. Die afwesigheid van metingsydigheid (bv. invariansie en ekwivalensie) moet dus empiries gedemonstreer word, in stede daarvan dat die afwesigheid daarvan eenvoudig aanvaar word (Van de Vijver & Tanzer, 2004). Die afwesigheid van hierdie informasie kan lei tot wydverspreide veralgemenings en ongetoetsde aannames wat die Sielkunde professie ernstige skade kan berokken. Die meetings-invariansie en -ekwivalensie van 'n instrument oor groepe is 'n voorvereiste vir substantiewe kruis-groepvergelykings (Dunbar, Theron & Spangenberg, 2011). Dit is noodsaaklik om empiries te bepaal of die instrumente wat gebruik is vry is van kulturele-, taal, geslag-, ouderdom- en rasse-sydigheid, nie net omdat dit verbied word deur die Wet op Diensbillikheid 55 van 1998 nie, maar ook omdat dit in die belang van goeie vakmanskap is. Sydigheid is sistermatiese steurnisse wat die geldigheid van die kruis-groep (kulturele) vergelykings (Van de Vijver & Leung, 1997) bedreig. Hierdie steurnisse kan wees as gevolg van konstruk-, metode- en/of itemsydigheid. Gegewe die belangrikheid van die besluite wat geneem word gebaseer op die metings is dit noodsaaklik dat die inligting vergelykbaar oor die verskillende verwysingsgroepe is. Die studie het die stel metingsinvariansie en -ekwivalensie toetse wat deur Dunbar et al. (2011) gebruik om 'n reeks van navorsingsvrae te beantwoord. Daar is ondersoek gestel na die mate waartoe die BWSI multi-groep metingsmodel as invariant of ekwivalent beskou kan word, en die bron van variansie te bepaal as dit sou bestaan (Vandenberg & Lance, 2000). In die ondersoek na die metingsmodel passing van die BWSI, is daar ondersteuning gevind is vir die hipoteses dat die metingsmodel beide van die geslagsteekproewe goed pas. Steun is ook gevind vir die konfigurale en swak invariansie modelle.

Aangesien slegs beperkte steun vir metriese ekwivalensie gevind is, is ondersoek na die parsiële metriese invariansie en ekwivalensie ingestel. Die BWSI multi-groep metingsmodel het voldoen aan die vereistes van parsiële volledige invariansie en parsiële volle ekwivalensie, en die nie-invariante items is deur die proses geïdentifiseer. Die implikasies van die resultate word bespreek, beperkinge word aangedui en areas vir verdere navorsing word uitgelig.

ACKNOWLEDGEMENTS

This research project would not have been successful without the guidance from my supervisors, Prof Callie Theron and Dr Gina Görgens. I would like to thank them for their support, commitment and meticulous feedback at each step of the way. It has been an honour to have worked with such inspirational lecturers who opened up my mind to the world of psychometrics, are passionate about the subject, and act with true integrity.

I would like to thank Dr. Nicola Taylor (Jopie van Rooyen & Partners) for authorising and supporting this research, and for providing the data needed for this study.

I would like to thank my family and friends who have always been supportive and encouraging. A special thanks to Kyle Davis, Lee de Andrade and Sam Barbosa for always supporting me, inspiring me and believing in me.

Lastly, but certainly not least, I would like to acknowledge my parents. To my mother, Maria Davis, I thank her for all her support, encouragement and love. She taught me to be strong, work hard and persevere. To my late father, Michael Davis, I appreciate and admire his love and commitment to his family and incredible sense of humour. He will always be missed and never forgotten. I dedicate this thesis to them.

TABLE OF CONTENTS

CHAPTER 1 ..... **1**

INTRODUCTION AND OBJECTIVE OF THE STUDY **1**

1.1 INTRODUCTION 1

1.2 RESEARCH OBJECTIVE 7

CHAPTER 2 ..... **9**

LITERATURE REVIEW OF THE SOURCES OF WORK STRESS INVENTORY **9**

2.1 OVERVIEW OF THE SWSI 9

2.2 DEVELOPMENT OF THE SWSI 10

2.3 PSYCHOMETRIC PROPERTIES OF THE SWSI: THE DEVELOPMENT OF THE SWSI 11

2.3.1 VALIDITY 11

2.3.2 RELIABILITY 15

2.4 PSYCHOMETRIC PROPERTIES OF THE SWSI: INDEPENDENT RESEARCH 16

2.4.1 VALIDITY OF THREE SWSI SCALES 16

2.4.2 RELIABILITY 21

2.4.3 GENDER DIFFERENCES 22

CHAPTER 3 ..... **23**

BIAS AND MEASUREMENT INVARIANCE AND EQUIVALENCE **23**

3.1 MEASUREMENT 23

3.2 BIAS IN MEASUREMENT 25

3.2.1 CONSTRUCT BIAS 26

3.2.2 METHOD BIAS 27

3.2.3 ITEM BIAS 28

3.3 INVARIANCE OR EQUIVALENCE IN MEASUREMENT 30

3.3.1 EVALUATING MEASUREMENT INVARIANCE AND EQUIVALENCE 31

3.3.2 TAXONOMY OF MEASUREMENT INVARIANCE AND EQUIVALENCE 33

3.4 PARTIAL INVARIANCE AND PARTIAL EQUIVALENCE 36

3.5 RESEARCH QUESTIONS 40

CHAPTER 4 ..... **44**

RESEARCH METHODOLOGY AND PRELIMINARY DATA ANALYSIS **44**

4.1 RESEARCH HYPOTHESES 44

4.2 RESEARCH DESIGN 45

4.3 STATISTICAL HYPOTHESES 46

4.4 SAMPLE 49

4.5 STATISTICAL ANALYSES 51
4.5.1 PREPARATORY PROCEDURES 51
4.5.1.1 MODEL SPECIFICATION 52
4.5.1.2 MODEL IDENTIFICATION 53
4.5.1.3 TREATMENT OF MISSING VALUES 54
4.5.1.4 ITEM ANALYSIS 55
4.5.1.5 DIMENSIONALITY ANALYSIS 56
4.5.2 STRUCTURAL EQUATION MODELLING 58
4.5.2.1 VARIABLE TYPE 58
4.5.2.2 EVALUATION OF MULTIVARIATE NORMALITY 59
4.5.2.3 MEASUREMENT MODEL FIT 59
4.5.2.4 DISCRIMINANT VALIDITY 60
4.5.2.5 TESTING FOR MEASUREMENT INVARIANCE AND EQUIVALENCE 61
4.6 STATISTICAL POWER 67
CHAPTER 5 69
RESULTS 69
5.1 INTRODUCTION 69
5.2 MISSING VALUES 69
5.3 ITEM ANALYSES 70
5.3.1 SUB-SCALE RELIABILITIES 71
5.3.2 ITEM STATISTICS 72
5.4 DIMENSIONALITY ANALYSES 73
5.4.1 DIMENSIONALITY ANALYSES RESULTS: MALE SAMPLE 74
5.4.2 DIMENSIONALITY ANALYSES RESULTS: FEMALE SAMPLE 85
5.5 CONCLUSIONS DERIVED FROM THE ITEM AND DIMENSIONALITY ANALYSES 96
5.6 MULTIVARIATE NORMALITY 97
5.7 EVALUATING THE SWSI SINGLE-GROUP MEASUREMENT MODEL FIT VIA CONFIRMATORY FACTOR ANALYSIS IN LISREL 99
5.7.1 SINGLE-GROUP MEASUREMENT MODEL FIT: MALE SAMPLE 99
5.7.2 SINGLE-GROUP MEASUREMENT MODEL FIT: FEMALE SAMPLE 113
5.7.3 SUMMARY OF SWSI SINGLE-GROUP MEASUREMENT MODEL FIT 123
5.8 DISCRIMINANT VALIDITY 123
5.8.1 AVERAGE VARIANCE EXTRACTED VERSUS SHARED VARIANCE: MALE SAMPLE 124
5.8.2 95% CONFIDENCE INTERVAL ESTIMATE: MALE SAMPLE 127

5.8.3 SUMMARY OF DISCRIMINANT VALIDITY: MALE SAMPLE 128
5.8.4 AVERAGE VARIANCE EXTRACTED VERSUS SHARED VARIANCE: FEMALE SAMPLE 128
5.8.5 95% CONFIDENCE INTERVAL ESTIMATE: FEMALE SAMPLE 131
5.8.6 SUMMARY OF DISCRIMINANT VALIDITY: FEMALE SAMPLE 132
5.9 CONFIGURAL INVARIANCE 132
5.10 WEAK INVARIANCE 136
5.11 METRIC EQUIVALENCE 139
5.11.1 DECISION ON THE RESULTS OF WEAK INVARIANCE AND METRIC EQUIVALENCE 141
5.12 PARTIAL METRIC EQUIVALENCE 141
5.12.1 DECISION ON THE RESULTS OF PARTIAL METRIC EQUIVALENCE 150
5.13 STRONG INVARIANCE 151
5.13.1 MEASUREMENT MODEL FIT INDICES 153
5.13.2 DECISION ON THE SUCCESS OF STRONG INVARIANCE 154
5.14 SCALAR EQUIVALENCE 155
5.14.1 DECISION ON THE RESULTS OF STRONG INVARIANCE AND SCALAR EQUIVALENCE .. 157
5.15 PARTIAL SCALAR EQUIVALENCE 157
5.15.1 DECISION ON THE RESULTS OF PARTIAL SCALAR EQUIVALENCE 167
5.16 STRICT INVARIANCE 168
5.16.1 MEASUREMENT MODEL FIT INDICES 170
5.16.2 DECISION ON THE SUCCESS OF STRICT INVARIANCE 171
5.17 CONDITIONAL PROBABILITY EQUIVALENCE 172
5.17.1 DECISION ON THE RESULTS OF STRICT INVARIANCE AND CONDITIONAL PROBABILITY EQUIVALENCE 173
5.18 PARTIAL CONDITIONAL PROBABILITY EQUIVALENCE 174
5.18.1 DECISION ON THE RESULTS OF PARTIAL CONDITIONAL PROBABILITY EQUIVALENCE 179
5.19 COMPLETE INVARIANCE 179
5.19.1 MEASUREMENT MODEL FIT INDICES 182
5.19.2 DECISION ON THE SUCCESS OF COMPLETE INVARIANCE 183
5.20 FULL EQUIVALENCE 183
5.20.1 DECISION ON THE RESULTS OF COMPLETE INVARIANCE AND FULL EQUIVALENCE ... 185
5.21 PARTIAL FULL EQUIVALENCE 186
5.21.1 DECISION ON THE SUCCESS OF PARTIAL FULL EQUIVALENCE 192
CHAPTER 6 193
DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS 193

6.1 INTRODUCTION 193
6.2 DISCUSSION 197
6.3 LIMITATIONS OF THE STUDY 201
6.4 RECOMMENDATIONS FOR RESEARCHERS AND PRACTITIONERS 202
REFERENCES 206
APPENDICES ..... CD

LIST OF TABLES

		Page
TABLE 2.1.	CORRELATIONS BETWEEN THE SCALE SCORES OF THE EIGHT SOURCES OF WORK STRESS SCALES	12
TABLE 2.2	CORRELATIONS BETWEEN FACTOR SCORES AND SCALE SCORES OF THE EIGHT SOURCES OF STRESS SCALES OF THE SWSI (N=416)	13
TABLE 2.3	SUMMARY OF THE RASCH RATING SCALE ANALYSIS RESULTS	15
TABLE 2.4	CRONBACH ALPHA COEFFICIENTS FOR THE SCALES OF THE SWSI	16
TABLE 2.5	OBLIQUE ROTATED FACTOR PATTERN MATRIX OF THE WORKLOAD, AUTONOMY AND GENERAL WORK STRESS ITEMS	17
TABLE 2.6	ITEM LOCATION PARAMETERS AND INFIT STATISTICS FOR THE GENERAL WORK STRESS, WORKLOAD AND LACK OF AUTONOMY SCALES	19
TABLE 2.7	OBLIQUE FACTOR STRUCTURE MATRIX OF THE NINE ITEMS OF THE GWSS (PROMAX, K = 4)	20
TABLE 2.8	A SUMMARY OF THE MEANS, STANDARD DEVIATIONS AND RELIABILITY STATISTIC FOR THE SWSI	21
TABLE 3.1	DEGREES OF MEASUREMENT INVARIANCE	34
TABLE 3.2	DEGREES OF MEASUREMENT EQUIVALENCE	35
TABLE 4.1	DEGREES OF FREEDOM FOR THE MULTI-GROUP MEASUREMENT INVARIANCE MODELS	53
TABLE 4.2	POWER CALCULATIONS	68
TABLE 5.1	NUMBER OF MISSING VALUES ACROSS ITEMS	70
TABLE 5.2	RELIABILITY OF THE SWSI SUB-SCALES FOR THE MALE SAMPLE	71
TABLE 5.3	RELIABILITY OF THE SWSI SUB-SCALES FOR THE FEMALE SAMPLE	72
TABLE 5.4	FACTOR ANALYSIS RESULTS FOR THE SWSI SUB-SCALES: MALE SAMPLE	75
TABLE 5.5	ROTATED FACTOR STRUCTURE FOR TH GENERAL WORK STRESS SUB-SCALE	76
TABLE 5.6	ROTATED FACTOR STRUCTURE FOR THE ROLE AMBIGUITY SUB-SCALE	77
TABLE 5.7	FACTOR MATRIX WHEN FORCING THE EXTRACTION OF A SINGLE FACTOR (ROLE AMBIGUITY)	78
TABLE 5.8	ROTATED FACTOR STRUCTURE FOR THE RELATIONSHIPS SUB-SCALE	79
TABLE 5.9	ROTATED FACTOR STRUCTURE FOR THE TOOLS AND EQUIPMENT SUB-SCALE	79
TABLE 5.10	ROTATED FACTOR STRUCTURE FOR THE CAREER ADVANCEMENT SUB-SCALE	80
TABLE 5.11	ROTATED FACTOR STRUCTURE FOR THE JOB SECURITY SUB-SCALE	81
TABLE 5.12	ROTATED FACTOR STRUCTURE FOR THE LACK OF AUTONOMY SUB-SCALE	82

TABLE 5.13 ROTATED FACTOR STRUCTURE FOR THE WORK/HOME INTERFACE SUB-SCALE 83
TABLE 5.14 FACTOR MATRIX WHEN FORCING THE EXTRACTION OF A SINGLE FACTOR (WORK/HOME INTERFACE) 83
TABLE 5.15 ROTATED FACTOR STRUCTURE FOR THE WORKLOAD SUB-SCALE 84
TABLE 5.16 PATTERN MATRIX WHEN FORCING THE EXTRACTION OF TWO FACTORS (WORKLOAD) 85
TABLE 5.17 FACTOR ANALYSIS RESULTS FOR THE SWSI SUB-SCALES: FEMALE SAMPLE 86
TABLE 5.18 ROTATED FACTOR STRUCTURE FOR THE GENERAL WORK STRESS SUB-SCALE 87
TABLE 5.19 ROTATED FACTOR STRUCTURE FOR THE ROLE AMBIGUITY SUB-SCALE 88
TABLE 5.20 FACTOR MATRIX WHEN FORCING THE EXTRACTION OF A SINGLE FACTOR (ROLE AMBIGUITY) 89
TABLE 5.21 ROTATED FACTOR STRUCTURE FOR THE RELATIONSHIPS SUB-SCALE 90
TABLE 5.22 ROTATED FACTOR STRUCTURE FOR THE TOOLS AND EQUIPMENT SUB-SCALE 90
TABLE 5.23 ROTATED FACTOR STRUCTURE FOR THE CAREER ADVANCEMENT SUB-SCALE 91
TABLE 5.24 ROTATED FACTOR STRUCTURE FOR THE JOB SECURITY SUB-SCALE 92
TABLE 5.25 PATTERN MATRIX WHEN FORCING THE EXTRACTION OF TWO FACTORS (JOB SECURITY) 92
TABLE 5.26 ROTATED FACTOR STRUCTURE FOR THE LACK OF AUTONOMY SUB-SCALE 93
TABLE 5.27 ROTATED FACTOR STRUCTURE FOR THE WORK/HOME INTERFACE SUB-SCALE 94
TABLE 5.28 PATTERN MATRIX WHEN FORCING THE EXTRACTION OF TWO FACTORS (WORK/HOME INTERFACE) 94
TABLE 5.29 ROTATED FACTOR STRUCTURE FOR THE WORKLOAD SUB-SCALE 95
TABLE 5.30 PATTERN MATRIX WHEN FORCING THE EXTRACTION OF TWO FACTORS (WORKLOAD) 96
TABLE 5.31 TESTS OF MULTIVARIATE NORMALITY FOR CONTINUOUS VARIABLES: BEFORE NORMALISATION 98
TABLE 5.32 TESTS OF MULITVARIATE NORMALITY FOR CONTINUOUS VARIABLES: AFTER NORMALISATION 98
TABLE 5.33 GOODNESS OF FIT STATISTICS FOR THE SWSI MEASUREMENT MODEL: MALE SAMPLE 101
TABLE 5.34 SWSI MEASUREMENT MODEL UNSTANDARDISED LAMBDA-X MATRIX (MALE SAMPLE) 109
TABLE 5.35 SWSI MEASUREMENT MODEL COMPLETELY STANDARDISED SOLUTION LAMBDA-X MATRIX (MALE SAMPLE) 110

TABLE 5.36 SWSI MEASUREMENT MODEL SQUARED MULTIPLE CORRELATIONS FOR X-VARIABLES (MALE SAMPLE) 112
TABLE 5.37 SWSI MEASUREMENT MODEL COMPLETELY STANDARDISED THETA-DELTA MATRIX (MALE SAMPLE) 112
TABLE 5.38 SWSI MEASUREMENT MODEL COMPLETELY STANDARDISED PHI MATRIX (MALE SAMPLE) 113
TABLE 5.39 GOODNESS OF FIT STATISTICS FOR THE SWSI MEASUREMENT MODEL: FEMALE SAMPLE 115
TABLE 5.40 SWSI MEASUREMENT MODEL UNSTANDARDISED LAMBDA-X MATRIX (FEMALE SAMPLE) 120
TABLE 5.41 SWSI MEASUREMENT MODEL COMPLETELY STANDARDISED SOLUTION LAMBDA-X MATRIX (FEMALE SAMPLE) 121
TABLE 5.42 SWSI MEASUREMENT MODEL SQUARED MULTIPLE SQUARED CORRELATIONS FOR X-VARIABLE (FEMALE SAMPLE) 122
TABLE 5.43 SWSI MEASUREMENT MODEL COMPLETELY STANDARDISED THETA-DELTA MATRIX (FEMALE SAMPLE) 122
TABLE 5.44 SWSI MEASUREMENT MODEL COMPLETELY STANDARDISED PHI MATRIX (FEMALE SAMPLE) 123
TABLE 5.45 AVERAGE VARIANCE EXTRACTED VERSUS SQUARED CORRELATION (MALE SAMPLE) 124
TABLE 5.46 95% CONFIDENCE INTERVAL ESTIMATE: MALE SAMPLE 127
TABLE 5.47 AVERAGE VARIANCE EXTRACTED VERSUS SQUARED CORRELATION (FEMALE SAMPLE) 128
TABLE 5.48 95% COFIDENCE INTERVAL ESTIMATE: FEMALE SAMPLE 131
TABLE 5.49 GOODNESS OF FIT STATISTICS FOR THE MULTI-GROUP SWSI CONFIGURAL INVARIANCE MEASUREMENT MODEL 134
TABLE 5.50 GOODNESS OF FIT STATISTICS FOR THE SWSI MULTI-GROUP WEAK INVARIANCE MEASUREMENT MODEL 137
TABLE 5.51 STATISTICAL SIGNIFICANCE OF THE SCALED CHI-SQUARED DIFFERENCE STATISTIC: A TEST OF METRIC EQUIVALENCE 139
TABLE 5.52 PRACTICAL SIGNIFICANCE OF THE CFI, GAMMA HAT AND MACDONALD DIFFERENCE STATISTIC: A TEST OF METRIC EQUIVALENCE 140
TABLE 5.53 STATISTICAL SIGNIFICANCE OF THE SCALED CHI-SQUARED DIFFERENCE STATISTIC: A TEST OF PARTIAL METRIC EQUIVALENCE (CONSTRUCTS) 143

TABLE 5.54 PRACTICAL SIGNIFICANCE OF THE CFI, GAMMA HAT AND MACDONALD DIFFERENCE STATISTIC: A TEST OF PARTIAL METRIC EQUIVALENCE PER SCALE 144
TABLE 5.55 LAMBDA_X_DIFFERENCE 146
TABLE 5.56 STATISTICAL SIGNIFICANCE OF THE SCALED CHI-SQUARED DIFFERENCE STATISTIC: A TEST OF PARTIAL METRIC EQUIVALENCE (ITEMS) 149
TABLE 5.57 PRACTICAL SIGNIFICANCE OF THE CFI, GAMMA HAT AND MACDONALD DIFFERENCE STATISTIC: A TEST OF PARTIAL METRIC EQUIVALENCE PER ITEM 150
TABLE 5.58 GOODNESS OF FIT STATISTICS FOR THE SWSI STRONG INVARIANCE MEASUREMENT MODEL 153
TABLE 5.59 STATISTICAL SIGNIFICANCE OF THE SCALED CHI-SQUARED DIFFERENCE STATISTIC: A TEST OF SCALAR EQUIVALENCE 155
TABLE 5.60 PRACTICAL SIGNIFICANCE OF THE CFI, GAMMA HAT AND MACDONALD DIFFERENCE STATISTIC: A TEST OF SCALAR EQUIVALENCE 156
TABLE 5.61 TAU_X_DIFFERENCE 158
TABLE 5.62 STATISTICAL SIGNIFICANCE OF THE SCALED CHI-SQUARED DIFFERENCE STATISTIC: A TEST OF PARTIAL SCALAR EQUIVALENCE PER INTERCEPT 160
TABLE 5.63 PRACTICAL SIGNIFICANCE OF THE CFI, GAMMA HAT AND MACDONALD DIFFERENCE STATISTIC: A TEST OF PARTIAL SCALAR EQUIVALENCE PER ITEM INTERCEPT 165
TABLE 5.64 GOODNESS OF FIT STATISTICS FOR THE SWSI STRICT INVARIANCE MEASUREMENT MODEL 170
TABLE 5.65 STATISTICAL SIGNIFICANCE OF THE SCALED CHI-SQUARED DIFFERENCE STATISTIC: A TEST OF CONDITIONAL PROBABILITY EQUIVALENCE 172
TABLE 5.66 PRACTICAL SIGNIFICANCE OF THE CFI, GAMMA HAT AND MACDONALD DIFFERENCE STATISTIC: A TEST OF CONDITIONAL PROBABILITY EQUIVALENCE 173
TABLE 5.67 THETA_DELTA_DIFFERENCES 175
TABLE 5.68 STATISTICAL SIGNIFICANCE OF THE SCALED CHI-SQUARED DIFFERENCE STATISTIC: A TEST OF PARTIAL CONDITIONAL PROBABILITY EQUIVALENCE PER ERROR VARIANCE 177
TABLE 5.69 PRACTICAL SIGNIFICANCE OF THE CFI, GAMMA HAT AND MACDONALD DIFFERENCE STATISTIC: A TEST OF PARTIAL CONDITIONAL PROBABILITY EQUIVALENCE PER ERROR VARIANCE 178
TABLE 5.70 GOODNESS OF FIT STATISTICS FOR THE SWSI COMPLETE INVARIANCE MEASUREMENT MODEL 182

TABLE 5.71 STATISTICAL SIGNIFICANCE OF THE SCALED CHI-SQUARED DIFFERENCE STATISTIC: A TEST OF FULL EQUIVALENCE 184
TABLE 5.72 PRACTICAL SIGNIFICANCE OF THE CFI, GAMMA HAT AND MACDONALD DIFFERENCE STATISTIC: A TEST OF FULL EQUIVALENCE 185
TABLE 5.73 VARIANCE DIFFERENCES 186
TABLE 5.74 COVARIANCE DIFFERENCES 187
TABLE 5.75 STATISTICAL SIGNIFICANCE OF THE SCALED CHI-SQUARED DIFFERENCE STATISTIC: A TEST OF PARTIAL FULL EQUIVALENCE PER VARIANCE AND COVARIANCE 189
TABLE 5.76 PRACTICAL SIGNIFICANCE OF THE CFI, GAMMA HAT AND MACDONALD DIFFERENCE STATISTIC: A TEST OF PARTIAL FULL EQUIVALENCE PER VARIANCE AND COVARIANCE 191

LIST OF FIGURES

Page

FIGURE 5.1	REPRESENTATION OF THE FITTED SWSI MEASUREMENT MODEL: MALE SAMPLE	100
FIGURE 5.2	STEM-AND-LEAF PLOT OF STANDARDISED RESIDUALS FOR THE SWSI MEASUREMENT MODEL FOR THE MALE SAMPLE	105
FIGURE 5.3	Q-PLOT OF SWSI MEASUREMENT MODEL STANDARDISED RESIDUALS FOR THE MALE SAMPLE	107
FIGURE 5.4	REPRESENTATION OF THE FITTED SWSI MEASUREMENT MODEL: FEMALE SAMPLE	114
FIGURE 5.5	STEM-AND-LEAF PLOT OF STANDARDISED RESIDUALS FOR THE SWSI MEASUREMENT MODEL FOR THE FEMALE SAMPLE	117
FIGURE 5.6	Q-PLOT OF THE SWSI MEASUREMENT MODEL STANDARDISED RESIDUALS FOR THE FEMALE SAMPLE	118
FIGURE 5.7	REPRESENTATION OF THE FITTED MUTLI-GROUP SWSI CONFIGURAL INVARIANCE MEASUREMENT MODEL FOR THE MALE AND FEMALE SAMPLES RESPECTIVELY	133
FIGURE 5.8	REPRESENTATION OF THE FITTED MULTI-GROUP SWSI WEAK INVARIANCE MEASUREMENT MODEL FOR THE MALE AND FEMALE SAMPLE RESPECTIVELY	136
FIGURE 5.9	REPRESENTATION OF THE FITTED MUTLI-GROUP SWSI STRONG INVARIANCE MEASUREMENT MODEL FOR THE MALE AND FEMALE SAMPLE RESPECTIVELY	152
FIGURE 5.10	REPRESENTATION OF THE FITTED MULTI-GROUP SWSI STRICT INVARIANCE MEASUREMENT MODEL FOR THE MALE AND FEMALE SAMPLE RESPECTIVELY	169
FIGURE 5.11	REPRESENTATION OF THE FITTED MULTI-GROUP SWSI COMPLETE INVARIANCE MEASUREMENT MODEL FOR THE MALE AND FEMALE SAMPLE RESPECTIVELY	181

CHAPTER 1
INTRODUCTION AND OBJECTIVE OF THE STUDY

This chapter aims to justify the objective of this research study through a systematic reasoned argument. It is essentially argued that occupational stress assessments play an important role in ensuring individual wellbeing, satisfaction and involvement at work and in ensuring that organisations are satisfied with the level of attendance and work performance their employees demonstrate. The interpretations derived from, and use of, occupational stress assessments across gender groups could be complicated due to lack of measurement invariance and equivalence, thereby hindering the abovementioned objective. Observed scores from measurement instruments may only be meaningfully compared across gender groups when measurement invariance and equivalence[1] have been established.

1.1 INTRODUCTION

Organisations are man-made phenomena that exist for a definite reason and with a specific purpose (Theron, 2011). An organisation exists to combine and transform scarce factors of production into products and services that society value. To ensure that organisations serve society in a rational manner the value of the products and services provided to the market should exceed the value of the factors of production committed to the transformation process. The primary goal of an organisation consequently, in a capitalistic system, is the maximisation of profit.

The effectiveness and efficiency with which organisations combine and transform factors of production into products and services with economic utility as well as identify and commercialise white space business opportunities and remain competitive in a perpetually changing and evolving business landscape, depend to a significant degree on the performance of their employees. The task of the human resource function in organisations is to affect the work performance of working man to the advantage of the organisation in a manner that adds value to the organisation. The human resource function seeks to contribute towards organisational goals through the attainment and maintenance of a competent, motivated and healthy workforce, as well as the effective and efficient utilisation of such a workforce (Theron, 2011). Human capital is a fundamental prerequisite to achieve organisational excellence. The workforce, furthermore, acts as an important factor for achieving projected organisational objectives (Hamidi & Eivazi, 2010). Industrial psychologists

[1] Although the terms invariance and equivalence are often used interchangeably in the literature, a clear distinction in meaning will be made between these two terms in this study. The terms will be defined in the literature study.

attempt to positively influence and improve the performance of working man through a variety of human resource interventions.

The management of employee wellbeing/psychological health is one of the human resource interventions with which the human resource function pursues this objective. Having happy and satisfied employees is of little value to an organisation unless employees are also performing efficiently and productively. Likewise, having an efficient and productive organisation is of little value if this is achieved at the expense of employees' wellbeing. Therefore, it is imperative to concurrently focus on employee wellbeing and performance in order to recognise this practical reality (Cotton & Hart, 2003). The employees' health and wellbeing is important to consider when wanting to optimise the productivity and effectiveness of the workforce in order to achieve a competitive advantage which all organisations strive for. The objective of the management of employee wellbeing/psychological health is not only to minimise the incidence of work performance pathology amongst employees but to actively promote employee wellbeing (Hofmann & Tetrick, 2003). Traditionally the management of employee wellbeing was seen as a process that was aimed at the prevention, detection and treatment of performance pathology (Hofmann & Tetrick, 2003). The focus was on pathology and its prevention and treatment. Recently, however, it was recognised that the management of employee wellbeing needs to move beyond the mere prevention and treatment of performance pathology and also actively promote positive psychological health if employee wellness interventions really want to contribute to organisational performance (Hofmann & Tetrick, 2003). Employees will not find an organisation an attractive employer, nor will current employees excel and experience low turnover intentions, if working at such an organisation means little more than a low probability of becoming unwell. The challenge facing the human resource function is to ensure that working at the specific organisation is instrumental in living a fulfilling, worthwhile, positive life. Work takes up a significant proportion of people's lives. Work need not be a disagreeable, painful means of earning the income needed to live life after hours and over weekends. Work can, and should, offer working man the opportunity to also find meaning in work.

To succeed in the management of employee wellbeing/psychological health the behaviour of working man needs to be validly understood. Employee wellbeing/psychological health is not a random event, but it is complexly determined by a nomological network of latent variables characterising the employee and his/her working environment. Given the point made earlier that the emphasis should be on employee wellness rather than the mere absence of pathology, the focus in uncovering the determinants should be on the positive conditions that should be present to foster wellness. Nonetheless, employee wellness cannot be achieved without the absence of negative

factors that detract from wellness and promote pathology. Although the focus should be on identifying the positive conditions that promote meaningful work and employee wellness, attention unavoidably also needs to remain on factors that detract from wellness. In the field of psychology and industrial psychology, thorough investigations have been undertaken, and the literature reflects this trend, that occupational stress adversely affects an individual's psychological and physical health, as well as an organisation's productivity and effectiveness (e.g. Cooper, Dewe & O'Driscoll, 2001; Leka, Griffiths & Cox, 2003; Van der Doef & Maes, 1999).

Work occupies a major part of people's lives, in terms of both time spent at work as well as the importance thereof (McLean, 1979). When the demands and pressures from work exceed an individual's knowledge, abilities and resources thereby challenging their ability to cope, work-related stress may occur. Work stress is a major, world-wide challenge to employee and organisational health. It is imperative for organisations to be aware of and sensitive to negative factors in the workplace, such as occupational stress, that influence employees' health and have a significant effect on job satisfaction and performance (Hamidi & Eivazi, 2010).

Broad inconsistencies exist in the way that stress is defined and operationalised. This discrepancy is compounded by the broad application of the stress concept throughout different disciplines. For example, the concept of stress has been defined as either a stimulus-based model (stress as an "independent" variable), a response-based model (stress as a "dependent" variable), or as a "process" (Cooper et al., 2001). The objectives of the research and the intended action resulting from the findings will typically influence the approach that is taken. Psychological stress, the form of stress this research study will focus on, is conceptualised by Schlebusch (1998) as 'an interaction of several variables involving a particular relationship between a person and the environment, which is appraised by the person as taxing or exceeding coping resources and endangering well-being' (p. 266). Therefore, stress is defined as a transaction which is concerned with the dynamics of the psychological mechanisms of cognitive appraisal and coping that underpins a stressful encounter (Cooper et al., 2001). Stress, here, is embedded in an ongoing process that involves individuals transacting with their environments, making appraisals of those stimuli, and attempting to cope with the issues that arise with the resources possessed.

One widely cited model of occupational stress is Karasek's (1979) job demand-control (JDC) model. Karasek's (1979) JDC model is based on the premise that the interaction between job demands and job control create different psychosocial work experiences for the individual, depending on the respective degrees of job demands and job control, and is the key to explaining strain-related outcomes. These work experiences were categorized into four types of jobs, namely high-strain jobs

(high demands and low control), active jobs (high demands and high control), low-strain jobs (low demands and high control), and passive jobs (low demands and low control). Therefore, strain occurs when high job pressures are combined with a perceived inability to influence tasks and procedures at work (Cooper et al., 2001).

Stress can have several negative outcomes for individuals that can be divided into three categories: behavioural strains, physical strains, and psychological strains (Spector, 2003). Psychological strains include emotional states, attitudes, and intentions. Physical strains can be immediate short term physiological disturbances and somatic symptoms, or long term illness. Behavioural strains are reactions to stressful conditions that can be adaptive or maladaptive and can be indicators of wellbeing (Spector, 2003). Therefore, negative individual outcomes can range from burnout (Doyle & Hind, 1998), to job dissatisfaction (Beehr, 1995), and cardiovascular disease (Theorell & Karasek, 1996). Furthermore, employee wellbeing largely affects organisational wellbeing.

There is a clear indication that psychosocial features of organisations can affect employee health and wellbeing, which in turn can affect organisations directly through increased costs due to absence and health claims, and indirectly through employees' reduced effectiveness (Spector, 2003). From an organisational perspective, Beehr (1995) indicated that employee withdrawal (lateness, absenteeism, turnover, and psychological withdrawal) and reduced job performance are two major organisational outcomes of occupational stress.

Many find work life stressful, and seem to accept it as part of the necessary frustration of daily life. Even though stress may be unavoidable, how one perceives and manages it is important. Yet, this is based on the presumption that there is sufficient understanding of what employee's perceive to be taxing in their work environment.

A well-known adage is that that in order to manage performance, valid and reliable measures of performance are required. The same principle applies to employee wellness and more specifically occupational stress. Prevailing stress levels need to be monitored regularly if escalating stress levels are to be detected in time to prevent serious personal and organisational problems from developing. The Sources of Work Stress Inventory (SWSI) is an instrument developed in South Africa specifically for this purpose (De Bruin & Taylor, 2005). One way of assisting the individual in understanding their work stress is to identify those variables that they perceive to be taxing in their work environment. Occupational stress assessment is a method that assists in indicating a general level of stress and highlights the possible sources of stress at work. The decisions that are made on the basis of work stress information will have a substantial impact not only on individuals but also organisations.

Individuals can use occupational stress assessments in order to isolate problem areas in their work environment in order to address them. For organisations, identifying the general level of employee stress can contribute towards a thorough organisational diagnosis. In terms of identifying the possible sources of stress, pinpointing the problem area in the workplace can lead to planning and implementing interventions to improve employee wellbeing/psychological health, and ultimately performance.

The SWSI attaches a specific connotative meaning (Kerlinger & Lee, 2000) to the stress construct. Specific latent stress dimensions are distinguished in terms of the SWSI's constitutive definition of stress. Specific items have been designed to serve as effect indicators (Hair, Black, Babin, Anderson & Tatham, 2006) of these latent stress dimensions. This design intention is reflected in the scoring key of the SWSI. The constitutive definition of the stress construct in conjunction with the design intention underlying the SWSI implies a very specific measurement model. A critical question is whether the measurement model reflecting the design intentions of the developers of the SWSI fits data obtained from the instrument at least reasonably well. Without credible psychometric evidence on the construct validity of the SWSI, its use in the management of employee wellness will not be warranted. Evidence on the psychometric integrity of the SWSI is reported in the literature (De Bruin & Taylor, 2005) and in the test manual (De Bruin & Taylor, 2006a). Quite sophisticated psychometric analyses have been performed on the SWSI. Despite this, however, none of the studies on the psychometric integrity of the SWSI have evaluated the fit, through confirmatory factor analytic procedures, of the (single-group) measurement model implied by the design intentions of the developers.

If reasonable single-group measurement model fit along with significant ($p < .05$) and reasonably high completely standardised factor loadings (at least .71 or higher; Hair et al., 2006) would be found when fitting the measurement model to a gender diverse sample, and when fitting the model to each gender group separately, it would permit the within gender group use of the SWSI to measure the stress construct as constitutively defined. Cross-gender group comparisons would, however, thereby not as yet be allowed. A further critical question that needs to be answered first is whether the measurement model parameters can be assumed to be the same across gender groups in South Africa?

The SWSI was developed using the JDC model as a foundation. However, the role of gender on the JDC model has not yet been fully explored (De Bruin & Taylor, 2006b). Van der Doef and Maes (1999) suggested that men and women may react differently to the effects of high-strain work, with men appearing more vulnerable to the negative effects of high job demands and low job control.

Furthermore, Vermeulen and Mustard (2000) concluded from their research that workplace characteristics, such as job demands and job control, may have a greater impact on the psychological wellbeing of men compared with women. This has prompted references to the JDC model as a "male model" (Johnson & Hall, 1988). In this regard, it is important to empirically determine whether the relations of job strain with job demands and job control are the same for important demographic groups, such as men and women. Measurement invariance and equivalence are necessary prerequisites[2] to empirically examine the structural invariance of the relationships across genders.

The inappropriate use of occupational stress assessments across genders can seriously jeopardise the extent to which occupational stress assessments, and the decisions based on them, achieve their intended objectives. New demands have been placed on psychological tests and practitioners that use these tests due to the changes in legislation (Patterson & Uys, 2005). Since 1994, stronger demands have been placed on the gender appropriateness of psychological tests, as outlined in the Employment Equity Act 55 of 1998. This places pressure on practitioners, test developers and test distributors to produce sophisticated scientific evidence that the instruments used in South Africa are psychometrically appropriate for, and relevant to, the South African context. Consequently, this creates the challenge to demonstrate that the measurement model underlying each test is transferable across groups.

Establishing measurement invariance and equivalence - ensuring that decision making is not based on two separate measurement models - is important and very relevant for South Africa. It is essential to establish whether the psychometric tools used in South Africa do not display group-related measurement bias and to ultimately minimise systematic error in as far as is achievable in a measure. In order to avoid making widespread generalisations and untested assumptions which will eventually do a disservice to the field of industrial psychology / psychology, the absence of measurement bias (i.e. invariance and equivalence) should be demonstrated instead of simply being assumed (Van de Vijver & Tanzer, 2004). Furthermore, Dunbar, Theron and Spangenberg (2011) indicate that establishing the measurement invariance and equivalence of an instrument across groups should be a prerequisite to conducting substantive cross-group comparisons. Therefore, it is imperative to empirically ascertain whether the instruments that are used are free of cultural, language, gender, age and racial bias, not only because it is prohibited by the Employment Equity Act 55 of 1998, but also in the interest of good workmanship.

[2] Specifically weak invariance and metric equivalence will have to be shown. These concepts will be defined in the literature study.

Equivalent numbers of occupational stress factors as well as equivalent pattern of factor loadings (i.e., configural invariance) is a necessary but not sufficient prerequisite to ensure that observed scores mean the same thing in terms of the underlying latent variable across gender groups. The magnitude of measurement model parameters could still differ across gender groups, even though the number of latent occupational stress dimensions and the pattern of factor loadings might be the same across gender groups. This difference would affect observed score interpretation. In order to compare observed test scores between gender groups and for meaningful inferences to be made with confidence, at least equal probability measurement equivalence needs to be established (Dunbar et al., 2011). Stated differently, in order to compare observed test scores between gender groups and for meaningful inferences to be made with confidence it needs to be demonstrated that the regression of item scores on the latent dimensions that they reflect do not differ in terms of intercept, slope or error variance across genders. In a nutshell, to compare observed test scores between gender groups, it needs to be shown that the SWSI measures are not gender biased (Foxcroft & Roodt, 2005; Theron, 2007).

The inferences made from the measuring instrument will be weak and questionable without sufficient evidence of measurement invariance and equivalence. Dunbar et al. (2011) indicated levels of invariance and equivalence that must be met before direct comparisons between different groups can be made. A variety of techniques exist that can be used to assess measurement invariance and equivalence, but there seems to be a consensus that multi-group confirmatory factor analysis, originally proposed by Jöreskog and Sörbom (1996a) and now commercially available through LISREL and other structural equation modelling software, represents one of the most accessible ways of testing cross-group comparisons of measurement instruments (Byrne, Shavelson & Muthén, 1989; Steenkamp & Baumgartner, 1998).

1.2 RESEARCH OBJECTIVE

This research study aims to evaluate the measurement invariance and equivalence of a South African occupational stress measure across gender groups. As highlighted above, appropriate work stress assessment affects both individual and organisational wellbeing. All previous studies based on the JDC model, that combined the data of men and women or that compared the derived scores of men and women, proceeded on the assumption that the scales are perceived in the same way and have the same meaning for the two groups (De Bruin & Taylor, 2006b). This is deemed inappropriate for the reasons explained above. It should be specified that this study does not aim to investigate gender definitions of occupational stress and resultant bias effects. The study solely aims to evaluate

the measurement invariance and equivalence of an occupational stress instrument, namely the Sources of Work Stress Inventory (SWSI), across gender groups in South Africa.

The SWSI authors attached a specific connotative meaning to the occupational stress latent variable. The connotative meaning of the occupational stress latent variable is set during the conceptualisation phase of the instrument development process since the manner in which a construct is used in an argument cannot be divorced from the meaning afforded to the construct. The connotative meaning of constructs firstly arises from the internal structure of the construct. The connotative meaning in addition arises from the manner in which the construct is embedded in a larger nomological network of latent variables. Specific latent work stress dimensions are distinguished in terms of the connotative meaning that the SWSI attaches to the occupational stress construct. Specific items have been designed to serve as effect indicators (Hair et al., 2006) of these latent work stress dimensions. This design intention is reflected in the scoring key of the SWSI.

A very specific measurement model is therefore implied by the design intentions of the developers of the SWSI and by the scoring key of the instrument. Critical questions in this study are whether [a] the single-group measurement model reflecting the design intentions of the developers fits data obtained from a gender diverse sample at least closely, [b] whether a multi-group measurement model reflecting the design intentions of the developers fits the data obtained from two separate gender samples at least closely and, if so, [c] whether the measurement model parameters differ across the two gender samples.

The objective of the research is to evaluate the fit of the single- and multi-group measurement model of the SWSI on a South African sample via confirmatory factor analysis (CFA) and to determine whether significant differences in measurement model parameters exist between samples of men and women.

CHAPTER 2
LITERATURE REVIEW OF THE SOURCES OF WORK STRESS INVENTORY

This chapter focuses on the SWSI. Existing literature is reviewed, providing an overview of the instrument. Furthermore, the processes followed in the development of the SWSI is outlined, the success with which the SWSI measures occupational stress as it is constitutively defined is evaluated, and empirical evidence is lead in support of the argument that the SWSI is a reliable and valid measure of occupational stress as it is constitutively defined, is presented.

2.1 OVERVIEW OF THE SWSI

The Sources of Work Stress Inventory (SWSI; De Bruin & Taylor, 2005) is an instrument developed to provide a measure of occupational stress that will not only reveal a general level of stress, but also focus on possible triggers or sources of stress (De Bruin & Taylor, 2005). The information obtained through the questionnaire will identify the sources of stress in order to address the areas of concern, which could eventually lead to a more appropriate working environment. The SWSI comprises of two segments: the General Work Stress scale and the Sources of Work Stress scales. The questionnaire consists of 59 items and takes approximately 20 - 30 minutes to complete (De Bruin & Taylor, 2005).

Forming the first section of the questionnaire, the General Work Stress scale is a brief self-report measure of an individual's overall level of subjectively experienced or "felt" work-related stress (De Bruin, 2006). The General Work Stress scale assists in measuring the degree to which individuals appraise their working environments as stressful. It therefore determines the extent to which work itself is a source of stress for the individual. The 9-item General Work Stress scale requests respondents to respond on a five-point Likert-type scale, indicating how often they experienced a certain feeling. The response categories include (1) Never, (2) Rarely, (3) Sometimes, (4) Often, and (5) Always.

The second part of the questionnaire, the Sources of Work Stress scale, assesses eight potential sources of work stress, namely: role ambiguity, relationships, tools and equipment, job security, career advancement, lack of autonomy, work/home interface, and workload. The 50-item Sources of Work Stress scale aims to assess the aspects of work that may cause stress. Respondents are required to answer on a five-point Likert-type scale, where they are required to report on the extent

to which each source of stress contributes to their level of stress at work. The response categories range from: (1) none at all, (2) very little, (3) some, (4) quite a lot, to (5) very much.

2.2 DEVELOPMENT OF THE SWSI

As part of the development of the SWSI, sources of stress were identified from the current stress literature. In addition, interviews were also conducted with various staff members at a university in order to identify aspects in the workplace which are perceived to cause stress. For each source of stress, items were drawn up which were then subjected to an item-sort with the use of index cards. Judges[3] were required to sort the cards with reference to predefined definitions of the various sources of stress. Ambiguous items or items that were not easily understood were modified, removed, or more clearly defined. Following the administration of the SWSI to 464 employees in an academic setting as well as extensive statistical analysis (factor analysis, item analysis and scale analysis), the authors arrived at a more parsimonious structure for the SWSI. Based on these results, the final instrument included the 9-item General Work Stress scale and 50 items measuring eight different sources of stress. These eight sources of work stress were conceptualised as follows (De Bruin & Taylor, 2005):

- *Role Ambiguity* refers to the amount of stress that an individual experiences as a result of continuous change and unclear requirements regarding the expectations, duties and constraints that define the individual's job.
- *Relationships* refer to the impact that poor interpersonal relationships with colleagues and superiors have on the level of stress experienced by the individual. It also refers to being subjected to interpersonal abuse.
- *Tools and Equipment* refers to stress experiences caused by a lack of appropriate tools and equipment required to do a job properly and includes working with inappropriate, broken and complex machinery.
- *Career Advancement* refers to the amount of stress that an individual experiences due to the perception that the individual has regarding a lack of opportunity to further his or her career prospects within the organisation for which he or she works.
- *Job Security*, as a source of stress, refers to the uncertainty regarding an individual's future in the current workplace, which then leads to stress.
- *Lack of Autonomy* refers to the amount of stress that an individual experiences due to a lack of empowerment in the workplace. It could also be viewed as job control or job decision latitude as referred to in Karasek's (1979) model of occupational stress.

[3] The judges consisted of three intern psychologists and a qualified psychologist.

- *Work/Home Interface* is a source of stress as a result of a lack of social support (family and friends) and work-non-work additivity, which refers to spill-over and conflict with regard to stress within and outside the workplace.
- *Workload* refers to an individual's experience of stress as a result of the perception that they are unable to cope or be productive with the amount of work allocated to them (De Bruin & Taylor, 2005).

2.3 PSYCHOMETRIC PROPERTIES OF THE SWSI: THE DEVELOPMENT OF THE SWSI

In this subsection, the reported psychometric properties (validity and reliability), as per the study of the development process of the SWSI (De Bruin & Taylor, 2005) will be discussed. The information presented in this subsection is based on data gathered from employees of a South African university that was about to enter a restructuring process, hence the need to evaluate the levels of stress experienced by the personnel.

2.3.1 VALIDITY

According to the traditional conceptualisation of validity, an instrument's validity reflects whether the test adequately measures what it claims to measure (Murphy & Davidshofer, 2005). Strictly speaking, however, it is not the instrument *per se* that is valid / invalid but rather the inferences derived from the instrument. The extent to which the test scores satisfy their intended purpose refers to its validity. It is important to assess the validity of a test as a basis for making specific decisions. The validity, or correctness, of the inferences made about people's levels on the construct being measured from the test scores is a major concern in psychological testing (Murphy & Davidshofer, 2005). Validity of the SWSI, as reported in the development of the instrument (De Bruin & Taylor, 2005; De Bruin & Taylor, 2006a), was evaluated by factor-analysis as well as Rasch rating scale model-based item analysis.

The intercorrelations of the original 79 Sources of Stress items were subjected to an unrestricted maximum likelihood factor analysis. Initially a nine-factor solution was extracted. The original nine factors explained 62.02% of the variance in the correlation matrix. The factor pattern matrix showed that, for five of the scales, items formed clusters that corresponded very well with the proposed scoring key, therefore providing strong support for the factorial validity of the following scales: Role ambiguity, Relationships, Tools and Equipment, Work/Home interface and Workload (De Bruin & Taylor, 2005). The remaining scales had to be reviewed. Following the merging of factors (Bureaucracy and Autonomy), splitting of factors (Job Security and Career Advancement) and

elimination of a factor (Physical Environment) and the items that loaded on it, an eight-factor solution was obtained for the 71 remaining items (De Bruin & Taylor, 2005). This solution explained 62.36% of the variance in the correlation matrix. Following maximum-likelihood factor analysis with a Promax rotation (k = 4) on the 50-item sources of work stress scales, an eight-factor solution consistent with the structure of the SWSI resulted. All items had salient primary loadings on their posited factors, except for item 35, which obtained similar loadings on Career Advancement, and Lack of Autonomy (De Bruin & Taylor, 2006a). Overall, the results from the factor analysis of the Sources of Stress items showed good fit with the proposed structure for the SWSI.

An examination of the intercorrelation matrix of the scale scores of the eight-factor solution (Table 2.1) indicated that all the scales are significantly correlated ($p < .01$) and therefore the possible presence of a general higher-order stress factor (De Bruin & Taylor, 2006a). This supports the premise that the Sources of Stress scales all measure some degree of stress (De Bruin & Taylor, 2005). This is further confirmed by the strong multiple correlation between the linear composite of the eight Sources of Stress scales and the General Work Stress scale (R = .66).

Table 2.1

Correlations between the scale scores of the eight sources of work stress scales

Scale	RA	REL	TE	CA	JS	LA	WH	WL
Role Ambiguity (RA)	1.00							
Relationships (REL)	.55	1.00						
Tools & Equipment (TE)	.36	.54	1.00					
Career Advancement (CA)	.45	.56	.45	1.00				
Job Security (JS)	.40	.45	.34	.60	1.00			
Lack of Autonomy (LA)	.47	.65	.54	.67	.54	1.00		
Work/Home Interface (WH)	.34	.39	.31	.30	.30	.41	1.00	
Workload (WL)	.33	.31	.31	.35	.28	.41	.58	1.00

All correlations are significant at the .01 level

(Adapted from De Bruin & Taylor, 2006a, p. 14)

The results for the factor analysis of the General Work Stress items were also satisfactory. After the deletion of four items, the remaining 11 items all had strong loadings on the single General Work Stress factor and appeared to define a psychologically coherent construct (De Bruin & Taylor, 2005).

Construct validity refers to the extent to which a measuring instrument measures the theoretical construct it was designed to measure in accordance with its constitutive definition (Cronbach & Meehl, 1955, as cited in Kerlinger & Lee, 2000). The connotative meaning of constructs ascribes a

specific internal structure to the construct. The internal structure specifies the dimensions comprising the construct and the manner in which these dimensions are related. Individuals with higher scores on the Sources of Stress scales are assumed to perceive a greater number of stressors in their environments and to experience these stressors more intensely than individuals with lower scores (De Bruin & Taylor, 2005). Support for the construct validity of the eight Sources of Work Stress scales is shown in Table 2.2. The correlations between the factor and scales scores for each of the Sources of Work Stress scales were all at or above .95.

Table 2.2

Correlations between factor scores and scale scores of the eight sources of stress scales of the SWSI (n=416)

	Factor Scores							
Scale	**RA**	**REL**	**TE**	**CA**	**JS**	**LA**	**WH**	**WL**
Role Ambiguity (RA)	**.95**	.51	.30	.45	-.41	-.40	.32	.30
Relationships (REL)	.47	**.99**	.51	.50	-.45	-.55	.36	.19
Tools & Equipment (TE)	.27	.48	**.99**	.39	-.36	-.45	.29	.21
Career Advancement (CA)	.39	.46	.40	**.98**	-.62	-.54	.22	.26
Job Security (JS)	.35	.37	.32	.59	**-.99**	-.45	.26	.21
Lack of Autonomy (LA)	.39	.58	.49	.64	-.54	**-.96**	.37	.31
Work/Home Interface (WH)	.30	.30	.30	.26	-.28	-.34	**.99**	.50
Workload (WL)	.27	.18	.31	.32	-.26	-.36	.57	**.96**

All correlations are significant at the .01 level. Correlations between corresponding scale and factor scores are printed on the diagonal in boldface.

(Adapted from De Bruin & Taylor, 2006a, p. 15)

De Bruin and Taylor (2005) further evaluated the construct validity of the SWSI scales by means of a multiple regression analysis. The relationships between the Sources of Stress scales and the General Work Stress scale provide support for their construct validity. The zero-order correlations between the General Work Stress scale and the Sources of Stress scales (as per the SWSI technical manual) were: Role ambiguity (r = .50), Relationships (r = .39), Tools and equipment (r = .27), Job security (r = .37), Career advancement (r = .40), Lack of Autonomy (r = .47), Work/Home interface (r = .48) and Workload (r = .54) (De Bruin & Taylor, 2006a). Although each Source of Stress scale could contribute to the prediction of General Work Stress, only Workload (r_{partial} = .30) and Role ambiguity (r_{partial} = .29) had meaningful partial correlations with General Work Stress, therefore suggesting that these two scales are the best predictors of General Work Stress (De Bruin & Taylor, 2006a).

Furthermore, the items of the eight Sources of Stress scales and the General Work Stress scale were subjected to Rasch rating scale analyses, using the Winsteps software (De Bruin & Taylor, 2005). The

Rasch model allows for a formal assessment of fit between a one-parameter item response model and the data, giving it an advantage over traditional item analysis procedures. According to the rating scale model, the probability that an individual will endorse a particular category of a particular item is a function of the individual's standing on the latent trait that the item measures, the overall difficulty or endorsability of the item and the difficulty in making the step to the chosen category from the preceding category (Bond & Fox, 2001). In the Rasch model it is required that the discrimination parameter is equal across all items. The slope of the probability that an individual will endorse a particular category of a particular item on the latent trait is therefore the same across all items. Items that do not meet this requirement will not fit the Rasch model well (Wright & Masters, 1982).

The results of the Rasch item analyses are summarised in Table 2.3. In this table, the lowest and highest values, respectively, are given for the INFIT mean squares, the item difficulty parameters and the item-score correlations. Only items with INFIT < 1.40 were retained in the item analysis, but very few items had to be discarded. Evidence that all the scales could be regarded as essentially unidimensional was provided by the fit indices. Furthermore, the item-score correlations were high for all nine scales, indicating that all items could be regarded as strong indicators of their respective traits (De Bruin & Taylor, 2005). According to De Bruin and Taylor (2005) the results obtained from the Rasch rating scale analysis in this study provide further support for the construct validity of the scales.

Table 2.3[4]

Summary of the Rasch rating scale analysis results

Measure	RA	RL	CA	JS	BA	WH	WL	TE	GWS
INFIT low	.63	.76	.67	.59	.74	.91	.63	.84	.70
INFIT high	1.38	1.24	1.29	1.35	1.29	1.11	1.36	1.35	1.29
Item reliability	.95	.98	.89	.93	.92	.99	.91	.93	.95
Person reliability	.82	.84	.76	.85	.91	.77	.85	.81	.88
Difficulty low	-.79	-.85	-.41	-.36	-.61	-1.11	-.29	-.40	-.77
Difficulty high	.29	.87	.25	.69	.42	1.15	.37	.62	.51
Item-score *r* low	.67	.75	.78	.88	.65	.62	.73	.67	.67
Item-score *r* high	.77	.87	.86	.94	.78	.81	.83	.82	.83

Notes: RA = Role Ambiguity, RL = Relationships, CA = Career Advancement, JS = Job Security, BA = Bureaucracy/Autonomy, WH = Work/Home Interface, WL = Workload, TE = Tools and Equipment, GWS = General Work Stress

(Adapted from De Bruin & Taylor, 2005, p. 761)

2.3.2 RELIABILITY

According to Kerlinger and Lee (2000) reliability in a measuring instrument refers to the degree that a measure is free from random measurement error. Classical measurement theory views reliability in a more technical manner as the proportion of systematic observed score variance (Theron, 2011). The reliability, or consistency, of test scores plays an important role in determining whether a test can provide a valid measure of the target construct (Murphy & Davidshofer, 2005). Reliabilities reported in the SWSI development study (De Bruin & Taylor, 2005) focus on Rasch rating scale analyses and internal consistency reliability coefficients (Cronbach alpha).

The results of the Rasch rating scale analyses (summarised in Table 2.3), also displayed the person-separation reliabilities and the item-separation reliabilities. The item-separation reliabilities were generally satisfactory. This suggests that the difficulty order of the items would be expected to remain the same if the analyses were repeated with a different sample of participants and that the items were well separated in terms of their difficulty parameters. The person-separation reliability estimates, which are analogous to Cronbach's alpha coefficient, were also satisfactory, suggesting that the items succeeded in separating individuals with different standings on the respective latent traits and that the order of the individuals on the trait would be expected to remain the same if a different sample of items were to be administered (Bond & Fox, 2001). According to De Bruin and

[4] The INFIT measure indicates the extent to which data satisfies the requirements of the model, item reliability is used to verify the item hierarchy, person reliability is used to classify people, the difficulty indicates the item difficulty range, and item-score correlations indicates whether the items can be regarded as strong indicators of their respective traits (De Bruin & Taylor, 2005).

Taylor (2005) the results obtained from the Rasch rating scale analysis in this study supported the reliability of the scales.

The Cronbach's alpha coefficients for the various Sources of Stress scales and the General Work Stress Scale were also calculated. These coefficients are displayed in Table 2.4. For the first version, values ranged from .86 to .95 and for the second version, values ranged from .86 to .94. These values can be described as satisfactory (De Bruin & Taylor, 2005; De Bruin & Taylor, 2006a).

Table 2.4

Cronbach alpha coefficients for the scales of the SWSI

	Version 1 (N = 311)		Version 2 (N = 464)	
Scale	**No. of items**	**A**	**No. of items**	**α**
General Work Stress	11	.92	9	.91
Role Ambiguity	9	.89	7	.87
Relationships	11	.93	8	.94
Tools & Equipment	8	.91	5	.90
Career Advancement	5	.90	5	.89
Job Security	4	.93	4	.92
Lack of Autonomy	17	.95	7	.90
Work/Home Interface	7	.86	8	.86
Workload	9	.93	6	.88

(Adapted from De Bruin & Taylor, 2006a, p. 13)

2.4 PSYCHOMETRIC PROPERTIES OF THE SWSI: INDEPENDENT RESEARCH

In this subsection, the reported psychometric properties of the SWSI, as per three additional studies, will be discussed. Statistical analyses (validity and reliability) of three SWSI scales (General Work Stress scale, Lack of Autonomy scale and Workload scale) will be examined. It should, however, be noted that no validation studies have been done with other existing measures of work stress.

2.4.1 VALIDITY OF THREE SWSI SCALES

De Bruin and Taylor (2006b) conducted research focusing on (a) the structural and measurement equivalence of measures of the job demand-control (JDC) model of job strain for men and women (as operationalised by the Sources of Work Stress Inventory), (b) whether a common or separate regression equations, with the regression of General Work Stress on job demands and job control, should be used for men and women in the JDC model, and (c) the strain and buffer hypotheses associated with the JDC model. The instruments used in this study included three scales of the Sources of Work Stress Inventory, namely the General Work Stress scale, the Workload Scale (used to operationalise job demands) and the Lack of Autonomy scale (used to operationalise job control).

The items of the three scales were subjected to maximum likelihood factor analyses for men and women in order to examine the convergent and discriminant validity and to assess the equivalence of the measured constructs across gender. The results[5] of the analyses for men and women separately indicated similar patterns of high and low factor pattern coefficients across gender. De Bruin and Taylor (2006b) concluded that the three factors represented qualitatively similar constructs for men and women. Results of the combined group of men and women are presented in Table 2.5.

Table 2.5

Oblique rotated factor pattern matrix of the Workload, Autonomy and General Work Stress items

Item		Factor		
		Workload	**Autonomy**	**General Work Stress**
WL1	No time for hobbies	**.584**	.033	.123
WL2	Work quickly	**.616**	.039	.135
WL3	Take work home	**.823**	-.093	.005
WL4	Work over weekends	**.627**	.050	.000
WL5	Cut back on social life	**.665**	.058	.019
WL6	Too few hours in day	**.821**	-.092	.112
WL7	Receive work at fast pace	**.755**	.052	.089
LA1	Changes happen too slow	.163	**.613**	-.047
LA2	Rigid rules	.016	**.745**	-.032
LA3	Policies and procedures prevent proper work	.140	**.730**	-.051
LA4	Unable to be creative	.064	**.746**	.025
LA5	Others make decisions about me	.099	**.747**	-.034
LA6	Not consulted on changes that affect me	.121	**.671**	.035
LA7	Ask permission before doing anything	.123	**.740**	-.089
GS1	Wish for different job	-.106	.269	**.759**
GS2	Want to quit	-.074	.286	**.769**
GS3	Worry about waking up and going to work	.007	.231	**.706**
GS4	Difficult to sleep at night	.170	.084	**.567**
GS5	So stressed forget to do important tasks	.207	.064	**.536**
GS6	So stressed difficult to concentrate on tasks	.254	.058	**.578**
GS7	Spend a lot of time worrying about work	.247	-.024	**.464**
GS8	Feel cannot cope with work anymore	.366	-.003	**.544**
GS9	So stressed that you lose your temper	-.024	.220	**.528**

Note. Factor pattern coefficients > 0.40 are printed in boldface. (Adapted from De Bruin & Taylor, 2006b, p. 69)
WL = Workload
LA = Lack of autonomy
GS = General work stress

[5] The coefficients of congruence of the corresponding factors for men and women were as follows: General work stress = .97, job control = .98 and job demands = .96 (De Bruin & Taylor, 2006).

Each factor was well defined as indicated by each variable noticeably loading on the factor it was expected to define. The correlations between the three factors (General work stress and job control, r = -.50 for men and r = -.37 for women; General work stress and job demands, r = .53 for men and r = .48 for women, and job demands and job control, r = -.34 for men and women) suggested that the relation between general work stress and job control differs for men and women (De Bruin & Taylor, 2006b). Overall, the convergent and discriminant validity of the items of the General Work Stress scale, Lack of Autonomy scale and Workload scale was supported by the results of the factor analysis, thereby providing support for construct validity. Empirical evidence indicated that the three scales measure separate but correlated constructs. Furthermore, the results indicate that qualitatively similar constructs were measured for men and women.

The items of each scale were also subjected to item response theory analyses. Specifically, the fit between the items and the requirements of the Rasch rating scale model were examined (see Table 2.6). The mean of the infit mean squares for the General Work Stress scale was .99 (SD = .15), which is close to the expected value of 1.00. The infit mean squares for the individual items of this scale ranged from .78 to 1.22, revealing that all the items demonstrated satisfactory fit. Strong item-measure correlations were found for all the items, ranging from .66 to .80. For the Workload scale, the mean of the infit mean squares was .99 (SD = .16), providing an indication of a satisfactory overall fit. All the individual items for the Workload scale demonstrated satisfactory fit as the infit mean squares ranged from .76 to 1.23. All the item-measure correlations were strong and ranged from .71 to .79. For the Lack of Autonomy scale, the mean of the infit mean squares was .99 (SD = .07). All the items indicated satisfactory fit, as the infit mean squares range from .94 to 1.12. Strong item-measure correlations were found for all the items, ranging from .73 to .78. These results confirmed that for each of the three scales a single line of enquiry runs through the items and that it is suitable to combine the items to obtain a single score or measure (De Bruin & Taylor, 2006b).

De Bruin and Taylor (2006b) further investigated differential item functioning (DIF) by comparing the item location parameters of the three scales for men and women. Across the three scales (General Work Stress, Workload and Lack of Autonomy), statistically significant differences in the item location parameter were found for only five items (approximately 22%). The impact of the item bias was considered minimal and practically unsubstantial (De Bruin & Taylor, 2006b). Taking these results and those of the factor analysis into account, it appears safe to assume that each of the three scales could be considered sufficiently unidimensional and internally consistent to warrant the computation of a total score for each participant. On the basis of these highly satisfactory results, it was concluded that men and women perceived the general work stress, job demands and job

control measures in the same way and that comparable measures were obtained for the two groups (De Bruin & Taylor, 2006b).

Table 2.6

Item location parameters and infit statistics for the General Work Stress, Workload and Lack of Autonomy scales

Item Label	Item Location Parameter	Standard Error	Infit Mean Square	Infit *t*	Item-Measure Correlation
		General Work Stress			
GS7	-.55	.06	1.22	3.3	.69
GS9	.05	.07	1.17	2.6	.66
GS4	.03	.07	1.12	1.8	.74
GS8	.11	.07	1.05	.80	.74
GS3	.05	.07	.98	-.30	.76
GS5	.46	.07	.93	-1.1	.70
GS1	-.60	.06	.89	-1.8	.78
GS2	.11	.07	.79	-3.5	.80
GS6	.34	.07	.78	-3.7	.76
Mean	0	.06	.99	-.20	
SD	.36	0	.15	2.4	
		Workload (Job demands)			
WL1	.02	.05	1.23	3.3	.71
WL4	-.16	.05	1.20	3.0	.72
WL2	-.33	.05	1.08	1.3	.72
WL5	.18	.05	.98	-.30	.75
WL3	.47	.05	.94	-.80	.76
WL7	.07	.05	.83	-2.6	.77
WL6	-.25	.05	.76	-3.9	.79
Mean	0	.05	.99	0	
SD	.25	0	.16	2.5	
		Lack of Autonomy (Job control)			
LA1	-.10	.06	1.12	1.7	.73
LA7	-.22	.06	.98	-.2	.75
LA6	-.50	.06	1.07	1.0	.76
LA2	.27	.06	.99	-.1	.75
LA4	.30	.06	.95	-.7	.77
LA5	-.09	.06	.91	-1.3	.78
LA3	.33	.06	.94	-.9	.77
Mean	.00	.06	.00	-.1	
SD	.29	.00	.07	1.0	

(Adapted from De Bruin & Taylor, 2006b, p. 69)

De Bruin (2006) conducted a further study that aimed to clarify the dimensionality or factor structure of the General Work Stress scale (GWSS). The GWSS was designed to function as a

unidimensional scale of work stress. Across the two independent data sets[6], the finding of three dimensions or factors of felt work stress (i.e. a motivational factor, an affective factor, and a cognitive factor) provided the best fit to the observed data, which appears to run counter to the model on which the scale is based (De Bruin, 2006). However, the factor structure coefficients indicate that each item correlated moderately to strongly with each of the three factors (Table 2.7) and the correlations of the three factors ranged from .692 to .711, which point toward the presence of a general factor.

Table 2.7

Oblique factor structure matrix of the nine items of the GWSS (Promax, k = 4)

	Group 1			Group 2		
	Factor 1	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**
G1	.874	.565	.576	.798	.586	.592
G2	.872	.613	.655	.883	.581	.590
G3	.808	.607	.596	.705	.526	.573
G4	.628	.581	.809	.599	.537	.740
G5	.562	.819	.567	.575	.874	.584
G6	.628	.878	.631	.671	.759	.635
G7	.504	.531	.764	.564	.578	.824
G8	.609	.700	.694	.719	.632	.623
G9	.606	.498	.473	.519	.538	.540

Note: All factor structure coefficients > 0.30 are underlined.

(Adapted from De Bruin, 2006, p. 71)

Furthermore, second-order factor analyses with a hierarchical Schmid-Leiman (1957) transformation showed that responses to the items are dominated by a general factor and that, in comparison, the influence of the three group factors is relatively weak (De Bruin, 2006). The general factor accounted for 72.4% and 74.7% of the shared variance for Group 1 and Group 2, respectively. McDonald's coefficient omega was calculated for each group. This represents the square of the correlation between the total score and the general factor that underlies responses to the items (McDonald, 1999). Omega for the two groups was .831 and .833, respectively. The square root of omega for group 1 was .911 and .913 for group 2. This indicates the correlations between the total score and the general factor, and shows that the total score is very strongly correlated with the hypothetical domain of which the items are a subset (De Bruin, 2006). This provides strong support for the construct validity of the total score.

6 Group 1 were 475 employees at two higher education institution (202 men and 273 women), and Group 2 were 477 employees at a large South Africa chemical company (97 women, 292 men and 88 of unknown gender).

2.4.2 RELIABILITY

The study by De Bruin and Taylor (2006b) in which the General Work Stress scale, Workload scale and the Lack of Autonomy scale were used, fitted the item data to the Rasch rating scale model. The person-separation reliability for the three scales was .89, .84, and .86, respectively. These values can be considered satisfactory and suggest that the items of the three scales are able to distinguish between individuals with different standings on the respective latent traits, and that the order of the individuals on the trait would be expected to remain the same if a different sample of items were to be administered (Bond & Fox, 2001).

De Bruin (2006) examined the dimensionality of the General Work Stress scale, which established the reliability of the obtained scores for the measuring instrument across the two sample groups. The reliability of the obtained scores for the General Work Stress scale for Group 1 and Group 2, as estimated by Cronbach's alpha coefficient, were .89 and .88, respectively (De Bruin, 2006).

Görgens-Ekermans and Brand (2012) investigated the inter-relationship between emotional intelligence (EI), work stress and burnout in the nursing industry and determined whether emotional intelligence is a moderator in the occupational stress and burnout relationship. The occupational stress construct was measured by the SWSI. Görgens-Ekermans and Brand (2012) developed reliability statistics for the SWSI (see Table 2.8). The Cronbach's alpha for each scale ranged from .78 to .94. This measure of internal consistency can be regarded as satisfactory although they generally tended to be somewhat lower than the values originally obtained by De Bruin and Taylor (2005).

Table 2.8

A summary of the means, standard deviations and reliability statistic for the SWSI

SWSI dimensions	Means	Standard Deviations	N of Items	Cronbach's Alpha
Role ambiguity	14.03	5.08	7	.79
Relationships	16.69	7.76	8	.92
Tools and Equipment	1.45	5.08	5	.94
Career Advancement	11.54	5.40	5	.84
Job security	8.50	3.92	4	.85
Lack of Authority	15.5	6.13	7	.85
Work/Home interface	13.91	5.24	7	.79
Workload	15.71	5.94	7	.78
General Work Stress	16.13	5.87	9	.87

(Adapted from Brand, 2007, p. 65)

2.4.3 GENDER DIFFERENCES

De Bruin and Taylor (2006b) investigated the differences between the genders for three of the SWSI scales (General Work Stress, Workload and Lack of Autonomy). The results of the DIF analyses showed that the scales measured the same traits and functioned equivalently for men and women (De Bruin & Taylor, 2006b). The observed score means of the men and women with regards to the linear combination of *General work stress*, *Job demands* (as operationalised by the Workload scale) and *Job control* (as operationalised by the Lack of Autonomy scale) were compared, and no statistically significant multivariate differences were shown. Furthermore, no statistically significant mean differences between men and women with regards to the three variables were shown as a result of univariate analysis of variance[7] (De Bruin & Taylor, 2006b). On the basis of the item response theory analyses results and the DIF results, it was concluded that men and women perceived the *General work stress*, *Job demands*, and *Job control* measures in the same way and that comparable measures were obtained for the two groups (De Bruin & Taylor, 2006b).

Following further analyses, the results of De Bruin and Taylor's (2006b) study indicated that, from a statistical perspective, the assumption of a common regression equation, describing the regression of General Work Stress on the linear composite of job demands and job control, for men and women did not hold. The results of a moderated hierarchical multiple regression analysis showed that *Job control* had statistically significant different slopes for men and women. *Job control* appears to be more strongly related to *General work stress* for men than for women (De Bruin & Taylor, 2006b). The results indicate that men respond more adversely to low control than women. The finding of the additive effects of *Job demands* and *Job control* accounting for slightly more variance in *General work stress* for men (approximately 40%) than for women (approximately 34%), potentially implies that the use of a common regression equation may slightly underestimate the strength of the additive effects of job demands and job control for men, and slightly overestimate the effects for women (De Bruin & Taylor, 2006b).

[7] It is unclear as to why the researchers performed and reported results of ANOVA if MANOVA reported no gender differences with regards to General Work Stress, Workload (job demands) and Lack of Autonomy (job control).

CHAPTER 3
BIAS AND MEASUREMENT INVARIANCE AND EQUIVALENCE

This chapter aims to clarify the connotative meaning of measurement bias, measurement invariance and measurement equivalence and to critically review literature on the methodology of detecting measurement bias, measurement invariance and measurement equivalence with the purpose of describing and justifying a best practice procedure of investigating measurement invariance and equivalence. The research methodology utilised in this study will be presented in the subsequent section.

3.1 MEASUREMENT

In the most general sense, measurement is the process of assigning numbers to objects in such a way that the numbers accurately represent the specific objects (Murphy & Davidshofer, 2005). Vandenberg and Lance (2000) defined measurement more specifically as "the systematic assignment of numbers on variables to represent characteristics of persons, objects, or events" (p. 4). The classic definition of Stevens (1946) defines measurement from an Industrial Psychology perspective as psychological assessments that attempt to indirectly measure latent variables through observable indicators in which the latent variables express themselves (reflective indicators)[8]. By eliciting a sample of observable behaviour through a sample of stimuli (items), latent psychological variables (like stress) are indirectly measured. Numerals are, therefore, assigned to behavioural indicators of the characteristic (Murphy & Davidshofer, 2005).

The ideal is that variance in the observed scores will only reflect variance in the latent variable that the indicator is meant to represent. Practically speaking, this ideal will never be fully realised. Nonetheless it is essential to implement processes that attempt to achieve the ideal that variance in the observed measure only reflects variance in the (to be) measured characteristics of persons, objects or events. This involves controlling extraneous variables that could cause non-relevant variance in the observed test scores as well as ensuring that the rules, referred to by Stevens (1946), are applied consistently. Through the processes of item analysis and standardisation, extraneous causes of variance in observed test scores are controlled for. This involves identifying and removing (if necessary) test items that do not primarily reflect the latent variable of interest, as well as standardising the test procedure so that the test stimuli, test instructions and test material remain the same across test takers and test administrators. However, these procedures never fully succeed

[8] The possibility of formative indicators (Hair et al., 2006) is not considered here.

in controlling non-relevant systematic and random influences that cause responses to test stimuli to vary.

The information obtained from measurement assessments is used with the intention of making informed decisions. Measurement, therefore, provides information on one or more latent variables relevant to the decision. This allows the decision-maker to determine a course of action based on the observations of the variables of interest. The quality of the information, and hence also the decision, depends on the measuring instrument; poor measurement can lead to incorrect decisions and interventions. One of the primary concerns in Industrial Psychology in terms of measurement is to ensure that the instrument does provide the appropriate information in order to make effective decisions and be able to accurately and consistently predict future behaviour (Theron, 2011).

The assumption that the latent variables of interest are measured reliably, validly and without bias is implicit in this argument. It is assumed that it is permissible to make inferences about the latent variables of interest and that these inferences are equivalent for members of different groups. If the measures do not reflect the same latent variables (i.e., general level of stress and sources of occupational stress) in the same, reasonably accurate manner across genders, it would be highly questionable to provide feedback to male and female employees based on these dimension measures. Furthermore, providing a more suitable working environment for employees would be seriously complicated if the measures, obtained for men and women, would reflect different occupational stress constructs or if specific observed measures of the general level of stress dimension and the sources of occupational stress dimensions would not indicate the same standing on the underlying latent variables[9].

Therefore, the questions are whether valid inferences can be made for members of both gender groups about the general level of stress and sources of occupational stress constructs, as they are constitutively defined by the SWSI, and whether the nature of such inferences would be the same across both gender groups. The former relates to whether the manner in which responses of men and women to the SWSI items relate to the underlying latent variables is the same (i.e., whether the factor structure necessary to closely reproduce the inter-item correlation matrix is the same across genders). The latter relates to whether the slope, the intercept and the error variance of the regression of the observed item responses on the latent traits are the same across both gender

[9] It could be argued that if uniform and/or non-uniform gender bias would exist in specific SWSI items and the nature of the bias would be such that it results in observed scale score differences between the genders that cannot be explained in terms of the latent variable, this situation still need not unavoidably result in the inferences on occupational stress and/or psychological wellbeing being biased. This line of reasoning, however, presupposes the development of an actuarial inference rule that would include a gender main effect and/or gender x predictor interaction effects. The development of such actuarial inference rules would undeniably pose severe practical, technical and logistical challenges to the psychologist.

groups. A fundamental issue is the comparability of scores across different gender groups, therefore the comparability of the instrument across genders should be investigated.

The ability to meaningfully interpret latent mean scores across gender groups point towards an equivalent psychological meaning of scores across gender groups, which means it is free from bias, or that invariance and equivalence have been established (Vandenberg & Lance, 2000). This implies that measurement instruments should be subjected to a series of statistical tests in order to be validated for use across gender groups (Theron, 2011). Empirically demonstrating the psychometric properties of the measurement instrument is necessary in the investigation of the cross-group applicability of the instrument (Van de Vijver & Poortinga, 1997).

3.2 BIAS IN MEASUREMENT

Measurement bias[10] represents the all systematic factors that could account for variance in observed test scores that cannot be accounted for in terms of the latent variable of interest (Theron, 2011). With test takers responding to a sample of questions or test stimuli under standardised conditions, a specific latent variable or construct is indirectly measured. It is assumed that the responses would be largely governed by the construct of interest however this is not always the case. The response to the test stimulus set is also influenced by other non-relevant, systematic factors and non-systematic, random factors that play a role. The non-relevant systematic nuisance factors essentially refer to any systematic source of unique variance in the test scores that cannot be explained in terms of variance in the latent variable of interest (Theron, 2011). In the analysis of measurement bias the emphasis usually focuses on sources of systematic measurement error that are systematically related to (cultural, language, gender) group membership. Differences in test scores between gender groups therefore might be due to differences in the construct of interest or due to systematic biases in the way the different gender groups respond to the items of the measurement instrument. If differences in the way the different gender groups respond to the items of the measuring instrument result in differences in test scores that cannot be explained in terms of differences in the same latent trait, it indicates that the test is biased.

Measurement bias, although undesirable, should not purely be seen as a nuisance factor. If differences exist in the manner in which people from different groups (irrespective of whether it would be language, gender, cultural or age groups) respond to the same test stimuli, this is presumably because of one or more latent variables that systematically differ across the groups in

[10] Measurement bias should be clearly distinguished from predictive bias. Predictive bias can be said to exist if group status, either as a main effect and/or in interaction with a single or a composite predictor, explains variance in the criterion that is not explained by the single or composite predictor.

question (Theron, 2011). Exploring the reasons for measurement bias (or more broadly lack of measurement invariance) should therefore be encouraged as a way of gaining greater understanding of group differences. Furthermore, Cheung and Rensvold (2002) maintained that invariance studies should also enhance understanding of the manner in which the groups being compared differ in the manner in which they respond to test stimuli. Investigations of any form of measurement invariance and equivalence should be seen as a source of potentially interesting and valuable information about how different groups view the world (Donnelly, 2009).

Measurement bias can occur due to a number of reasons. Bias does not occur due to the inherent properties of the measuring instrument, but exists due to the characteristics and traits of the respondents in the different groups that utilise the instrument (Van de Vijver & Poortinga, 1997). Van de Vijver and Poortinga (1997) developed a taxonomy to describe different types of bias that should be identified prior to making cross-group comparisons. These sources of bias include construct bias, method bias, and item bias.

3.2.1 CONSTRUCT BIAS

When the construct measured by the instrument is not the same across groups, construct bias occurs (Van de Vijver & Leung, 1997). From a measurement perspective, construct bias is indicated when the observed scores do not reflect the same construct across groups. Evidence of construct bias is apparent when the number of underlying latent variables (or factors) that needs to be assumed to satisfactorily account for the covariance in the test responses of different groups, the degree to which the latent variables are inter-related, and/or the specific latent variables that underlie the observed responses to each test stimulus (or item), differ across groups. Stated more concisely, construct bias exists if the factor structure that is required to closely reproduce the observed covariance matrix differs across groups in terms of number of factors, correlation between factors and/or loading pattern. This interpretation does not as yet fully reflect the fact that the connotative meaning of constructs not only arises from the internal structure of the construct but also from the manner in which the construct is embedded in a larger nomological network of latent variables. Construct bias therefore also exists if the manner in which the construct is embedded in a larger nomological network of latent variables differs across groups. Construct bias therefore occurs when the nature of the structural model that needs to be assumed to satisfactorily account for the observed inter-item covariance matrix differs across groups.

It is important to understand why different (factor analytic and/or structural model) explanations would be required to account for the manner in which people from different groups respond to the, objectively speaking, same set of test stimuli. It essentially means that behaviours that serve as

denotations of a specific construct in one group do not do so in another group. Another potential source of construct bias is the inadequate domain sampling of the behaviours in different groups. A further cause may be due to an instrument not capturing complete coverage of a construct's sub-domain (Van de Vijver & Poortinga, 1997).

3.2.2 METHOD BIAS

When members of specific groups respond differently to the same set of test items, and the reason for the different responses cannot be explained in terms of the target factors being measured, but is rather caused by group-related variables, then method bias exists. Method bias rather serves to explain item (and possibly construct) bias (Theron, 2011) and is not an additional, unique form of bias that describes a unique aspect of the nature of the relationship between observed scores on indicator variables and the latent variables underlying the responses to the indicator variables. In terms of this line of thinking method bias provides explanations as to why construct bias and especially item bias occurs. Identification of the sources of method bias may result in the researcher avoiding the variance caused by it, in the results obtained. According to Van de Vijver and Rothman (2004), method bias includes sample bias, administration bias and instrument bias.

Sample bias relates to the lack of comparability of the samples on other factors than the construct being assessed for example, biographical and demographic variables (Byrne & Watkins, 2003). Ideally, the samples should be reasonably comparable in terms of these variables. Administration bias is attributed to differences in the method used to administer an instrument. For example, one group might have been guided through the practice items and the other group did not receive this practice (Byrne & Watkins, 2003). Interaction between the test administrator and the respondent may also be associated with such bias. This could be due to communication problems in the case of a test administrator's use of language, understanding of cultural norms or personal bias against the group being tested (Van de Vijver & Leung, 1997). However, it is assumed that the testing conditions would have been in accordance with the standardised conditions specified in the administration manual as trained professionals[11] are the custodians of the assessment process. Instrument bias derives from problems associated with the measurement instrument that cause unintended cross-group differences (Byrne & Watkins, 2003; Van de Vijver & Rothman, 2004).

The four most common sources of method bias include (i) differential social desirability (DSD), (ii) differential response style (DRS), (iii) differential stimulus familiarity (DSF), and (iv) group differences

[11] In South Africa, a trained professional is an individual that is registered with the Health Professionals Council of South Africa (Professional Board for Psychology) as being permitted to carry out psychological assessment. These registered individuals are required to complete academic, supervision and examination requirements prior to registration and consequent access to psychological tests/questionnaires (Health Professions Act, 1974).

on the latent variables that affect the response to test items (Berry, Poortinga, Segall & Dasan, 2002; Byrne & Watkins, 2003; Theron 2011). It is possible that members of one group, therefore, (a) tend to systematically respond in a more socially desirable manner to test stimuli than members of another group (i.e., DSD), (b) could be more (or less) familiar with the test stimuli than members of another group (i.e., DSF), (c) tend to favour certain response alternatives more (or less) than members of another group (i.e., DRS) or (d) tend to be systematically different to members of another group on (non-relevant) characteristics that are related to test responses (Theron, 2011).

Social desirability refers to the tendency to want to present a favourable impression of oneself when responding to questionnaire items in terms of prevailing norms (Edwards, 1957; Edwards, 1970). In stress questionnaires there may be the possibility that men may respond negatively to "female" stressors such as "work/home interface" and women may respond negatively to "male" stressors such as "job security"- not because of it not being a source of stress for them, but for fear of displaying stress from a gender stereotypical source of stress. This could manifest as socially desirable responding in each gender group.

Differential response styles can be related to acquiescence or extreme rating[12], which may be more prevalent in particular (cultural) groups. Such response styles may be driven by culture, and would threaten the validity of results. There is research that suggests that ERS may also be gender related. Johnson, O'Rouke, Chavez, Sudman, Warnecke and Lacey (1997) have provided some evidence to suggest that age and gender may be associated with response artifacts (e.g. ERS). More recently, Johnson, T.J., Kulesa, P., Cho, Y.I., & Shavitt, S. (2005) reported significant age and gender differences in ERS across 19 nations. Another main source of systematic nuisance variance would be a group's familiarity with a stimulus that is used to assess a particular domain. Typically differences in stimulus familiarity would occur if behavioural denotations are used as stimuli that are more common in one group than another, or that one group is more familiar with than another. With regards to typical performance measures/questionnaires that require degrees of preference to be indicated, perhaps through a Likert-type scale, differential stimulus familiarity may even occur in groups that have not been exposed to a Likert-type scale (Berry et al., 2002, Byrne & Watkins, 2003).

3.2.3 ITEM BIAS

Item bias refers to undesirable measurement artefacts at the item level (Van de Vijver & Leung, 1997). This is often referred to as differential item functioning (DIF). Item bias exists if group

[12] Acquiescence Response Style is also known as agreement bias, regardless of item content (Johnson, Kulesa, Cho, & Shavitt, 2005). Extreme Response Style is the tendency to use the extreme ends of a rating scale (Cheung & Rensveld, 2002).

membership explains variance in the observed item response (either as a main effect and/or in interaction with the latent variable being measured) that is not explained by the latent variable being measured (Theron, 2011).

From a different perspective, in terms of a rather stringent definition of item bias, item bias could be said to exist if the probability of achieving a specific observed score on the item would differ across groups for individuals with the same standing on the latent variable being measured (Theron, 2011). In terms of a slightly more lenient definition of item bias, item bias could be said to exist if the expected observed score on the item would differ across groups for individuals with the same standing on the latent variable being measured (Theron, 2011).

Item bias, in terms of the more lenient definition of item bias, would exist if the regression of the observed response on the latent variable being measured would differ across groups in terms of slope and/or intercept (Theron, 2011). The former situation is referred to as non-uniform bias (Van de Vijver & Leung, 1997) and would imply a group x latent variable interaction effect on the observed item response, whereas the latter situation is known as uniform bias and would imply a group main effect on the observed item response (Van de Vijver & Leung, 1997). In terms of the more stringent definition of item bias, item bias would exist if the regression of the observed response on the latent variable being measured would differ across groups in terms of slope and/or intercept and/or residual error variance (Theron, 2011).

Determining why the regression of the observed test response on the latent variable would differ across groups is important. The sources of method bias discussed earlier could provide these explanations. Additional sources of method bias that could explain item bias include poor translation of items, inadequate item formulation (using complex wording, double-negatives, idiomatic expressions), items that tap into other constructs and the appropriateness of item content for the target group (understanding of item content for the testing context).

According to De Beer (2004), item bias should be investigated and corrected during the process of instrument construction. To ensure group appropriate instruments, the process needs to include a phase of item analysis that has to include item bias analysis aimed at identifying and eliminating biased items. If removing inappropriate items or indicators results in a decrease in measurement bias, it may be deduced that any previously observed score differences were likely due to item bias and not inherent differences across groups in the construct of interest (Van de Vijver & Leung, 1997).

3.3 INVARIANCE OR EQUIVALENCE IN MEASUREMENT

Measurement bias, measurement invariance and measurement equivalence are critical concepts in cross-group assessment. Measurement bias, measurement invariance and measurement equivalence essentially refer to the same phenomenon albeit from two different perspectives. Measurement bias, referring to all nuisance factors that prevent the meaningful comparison of scores across groups by affecting either the factor structure or the intercept, slope or error variance of the regression of the indicator on the latent variable being measured, approaches error in measurement from the perspective of classical measurement theory and item response theory. Measurement invariance and measurement equivalence likewise focus on whether the factor structure differs across groups or whether the intercept, slope or error variance of the regression of the indicator on the latent variable being measured differs across groups. Measurement invariance and equivalence, however, approach these same questions from the perspective of structural equation modelling. Construct and item bias would manifest in differences in measurement model characteristics[13] across groups (Theron, 2011). These two forms of measurement bias are in essence defined from the perspective of structural equation modelling in terms of the manner in which the measurement model underlying the test differs across groups. Measurement invariance and equivalence (or rather the lack thereof) therefore presents a different perspective on systematic errors in measurement, but essentially refers to the same issues as that of construct and item bias. It can possibly be claimed that by viewing systematic errors in measurement from the perspective of structural equation modelling, a more detailed and more finely enunciated evaluation of measurement bias is obtained.

When viewed from the perspective of structural equation modelling, like when viewed from the perspective of classical measurement theory, method bias does not translate to unique problems with measurement model characteristics that are not already covered by the concepts of construct and item bias. Unlike construct and item bias, method bias cannot be defined in terms of unique differences in measurement model characteristics (Theron, 2011). Rather method bias embodies the plea of Berry et al. (2002) and Cheung and Rensvold (2002) that attempts should be made to understand construct and item bias (i.e., differences in measurement model characteristics) when it occurs.

As implied by Horn and McArdle (cited in Vandenberg & Lance, 2000), insufficient evidence indicating measurement invariance and equivalence will result in a lack of confidence with regards to

[13] Reference is made here of measurement model characteristics rather than measurement model parameters so as to accommodate differences in the number of latent variables and loading patterns and not only differences in Λ, $\boldsymbol{\phi}$ or $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$.

the scientific inferences drawn from measurement instruments. Differences between individuals and groups cannot be interpreted unambiguously in the absence of such evidence. Confirmation of invariance and equivalence indicates the absence of factors that challenge the validity of cross-group comparisons. Testing for, and establishing measurement, invariance and equivalence is therefore a logical and important prerequisite for conducting cross-group comparisons. Detecting the presence of invariance and equivalence will help guide the development of more appropriate and sound instruments, as well as place more confidence in the validity of test results and the comparability of scores across groups.

3.3.1 EVALUATING MEASUREMENT INVARIANCE AND EQUIVALENCE

Historically, the quality of psychological tests has been evaluated through the classical test theory (CTT) of true and error scores (Crocker & Algina, 1986; Nunnally & Bernstein, 1994). Vandenberg and Lance (2000) acknowledged that CTT provides valuable information regarding the reliability and validity of measurement instrument properties. However, simple reliability and validity studies tend to ignore the issue of invariant and equivalent models of measurement. The extent to which measurement instrument properties are transportable across populations reflects the main question in terms of measurement invariance and measurement equivalence. Vandenberg (2002) argued that a lack of measurement invariance and equivalence threatens the value of measurement instruments that are not directly addressable through the classical test theory approaches, such as the calculation of reliability coefficients. The CTT's primary concern is to what extent the measurement instrument (X) can be used as a representation of the latent variable of interest (ξ). CTT does not easily extend into tests that directly determine whether there is conceptual equivalence of the construct of interest (ξ) in each group, or equivalent associations (λ) between operationalisations (X) across groups, and the extent to which the measurement instrument (X) is influenced to the same degree and by the same unique factors (δ) across groups (Vandenberg & Lance, 2000). To this end, Vandenberg and Lance (2000) argued that investigating measurement invariance and equivalence is just as important as providing proof of the reliability and validity of measurement instruments.

Without appropriate empirical research evidence the question regarding the equivalence of the parameters of the measurement model underlying the SWSI across gender remains unanswered and therefore high on the research agenda. This is due to measurement invariance and measurement equivalence being a necessary prerequisite for any meaningful cross-gender comparisons. It should be acknowledged that measurement invariance and equivalence only recently started receiving increased research attention (Vandenberg & Lance, 2000; Vandenberg 2002). The relatively recent increase in investigations of measurement invariance and measurement equivalence is possibly due

to the relatively recent developments in the data analytical tools. In this study, measurement invariance and measurement equivalence will be evaluated according to a confirmatory factor analytical (CFA) framework. A number of specific aspects to the measurement invariance and measurement equivalence issues are readily testable within a CFA framework.

Vandenberg and Lance (2000) explained the relationship between observed scores on indicator variables, and the latent variables the indicator variables are meant to reflect, through the following mathematical equation:

$$\mathbf{X}^g = \boldsymbol{\tau}^g + \boldsymbol{\Lambda}^g\boldsymbol{\xi}^g + \boldsymbol{\delta}^g$$

$\mathbf{X}^g$ refers to the vector of items comprising the measuring instrument of the g^{th} group, $\boldsymbol{\tau}^g$ refers to the vector of regression intercepts, $\boldsymbol{\Lambda}^g$ refers to the matrix of the regression slopes relating the items (X^g) to the construct of interest ($\boldsymbol{\xi}^g$), and $\boldsymbol{\delta}^g$ refers to the vector of unique factors. This equation, however, fails to fully capture the measurement model as it does not identify the manner in which the latent variables are related and the manner in which the measurement error terms are related. Assuming that $E(\xi^g, \delta^g) = 0$, the covariance equation that follows from the above mentioned equation is (Vandenberg & Lance, 2000):

$$\boldsymbol{\Sigma}^g = \boldsymbol{\Lambda}^g_x\, \boldsymbol{\phi}^g\, \boldsymbol{\Lambda}^{g'}_x + \boldsymbol{\Theta}_\delta{}^g$$

Where $\boldsymbol{\Sigma}^g$ is the matrix of indicator variable variances and covariances in the g^{th} population group, $\boldsymbol{\Lambda}^g_x$ is the matrix of items factor loadings on $\boldsymbol{\xi}^g$, $\boldsymbol{\phi}^g$ contains the variances and covariances among the $\boldsymbol{\xi}^g$, and $\boldsymbol{\Theta}_\delta{}^g$ is the diagonal matrix of unique variances. This is the fundamental covariance equation for factor analysis that models the observed item covariances as a function of common ($\boldsymbol{\xi}^g$) and unique ($\boldsymbol{\delta}^g$) factors.

From the above mentioned equations it becomes clear that aspects related to the measurement invariance and equivalence issues are testable within a CFA framework. As stated by Vandenberg and Lance (2000) the equations imply the following as testable hypotheses relating to measurement equivalence:

- $\boldsymbol{\xi}^{g'} = \boldsymbol{\xi}^{g'}$, the items of the measuring instrument evokes the same conceptual framework in defining the construct of interest (ξ) in each group.
- The CFA model holds equivalently and assumes a common form across groups.
- $\boldsymbol{\Lambda}^g = \boldsymbol{\Lambda}^{g'}$, the regression slopes linking the measures (X) to the underlying construct of interest (ξ) are invariant across groups.

- $\boldsymbol{\tau}^{g'} = \boldsymbol{\tau}^{g'}$, the regression intercepts linking the measures (X) to the underlying construct of interest (ξ) are invariant across groups.

- $\boldsymbol{\Theta}_{\delta}^{g} = \boldsymbol{\Theta}_{\delta}^{g'}$, the unique variances for the measuring instrument are invariant across groups.

- $\boldsymbol{\phi}^{g} = \boldsymbol{\phi}^{g'}$, the variances and covariances among the latent variables are invariant across groups.

Establishing the measurement invariance and equivalence of an instrument across groups should be a prerequisite to conducting substantive cross-group comparisons. The conviction with which inferences can be drawn is reduced without supporting measurement invariance and equivalence evidence of the instrument (Horn & McArdle, 1992). Questions about using the specific instrument within heterogeneous groups (or across homogeneous groups) are raised if measurement invariance and measurement equivalence are not established for a measure such as the SWSI (Steenkamp & Baumgartner, 1998). This is due to the fact that, in the absence of evidence of measurement invariance and equivalence, findings of differences between individuals and groups cannot be interpreted unambiguously. Researchers (e.g., Lubke & Muthén, 2004; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000) have indicated that the lack of invariance and equivalence studies can be ascribed to various factors including (a) different terminology used for the different types of invariance and equivalence found in literature which causes confusion, (b) the complex and unfamiliar methodological procedures used to test for different types of invariance and equivalence and (c) the few guidelines provided to help determine whether a measure exhibits invariance and equivalence. This has led researchers to endeavour to clarify key invariance and equivalence issues and to propose best practices for establishing invariance and equivalence (e.g. Byrne & Watkins, 2003; Cheung & Rensvold, 2002; Vandenberg, 2002; Vandenberg & Lance, 2000).

Dunbar, et al. (2011) have proposed a taxonomy of measurement invariance and measurement equivalence which aspires to contribute to a convergence towards a uniform understanding of, and approach towards, invariance and equivalence research.

3.3.2 TAXONOMY OF MEASUREMENT INVARIANCE AND EQUIVALENCE

In measurement invariance and equivalence research, two sets of questions arise. Clearly differentiating between the two sets of questions will decrease the current semantic confusion (Dunbar et al., 2011). Determining whether a multi-group measurement model with none of its parameters constrained to be equal across groups or with equality constraints imposed on some of its parameters or with all its parameters constrained to be equal across groups fits the data obtained from two or more samples, refers to the first set of questions (Dunbar et al., 2011). In the second set of questions, two multi-group measurement models are compared. The question here asks whether

a specific multi-group measurement model with some of its parameters constrained to be equal across groups fits substantially poorer than a multi-group model with fewer of its parameters constrained to be equal across groups (Dunbar et al., 2011). To assist in separating these two sets of questions, Dunbar et al. (2011) proposed that the term "measurement invariance" be reserved to refer to the first set of questions. Five hierarchical levels of measurement invariance are distinguished in Table 3.1, which were first introduced by Meredith (1993). The five levels refer to multi-group measurement models with increasing constraints placed on the model that fits the data of two or more groups, thus clearly referring to the first set of questions (Dunbar et al., 2011). Table 3.1 presents the various forms of measurement invariance distinguished by Meredith (1993) and provides a definition of each form of invariance.

Table 3.1

Degrees of measurement invariance

Configural invariance	**Weak invariance**	**Strong invariance**	**Strict invariance**	**Complete invariance**
A multi-group measurement model in which the structure of the model is constrained to be the same across groups fits multi-group data.	A multi-group measurement model in which the structure of the model is constrained to be the same across groups and in which the factor loading matrix (Λ^X) is constrained to be the same across groups fits multi-group data.	A multi-group measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^X is constrained to be the same across groups and in which the vector of regression intercepts (τ^X) is constrained to be the same across groups fits multi-group data.	A multi-group measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^X is constrained to be the same across groups, in which τ^X is constrained to be the same across groups and in which the measurement error variance-covariance matrix (Θ_δ) is constrained to be the same across groups fits multi-group data.	A multi-group measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^X is constrained to be the same across groups, in which τ^X is constrained to be the same across groups, in which Θ_δ is constrained to be the same across groups and in which the latent variable variance-covariance matrix (Φ) is constrained to be the same across groups fits multi-group data.

(Dunbar et al., 2011, p. 11)

Dunbar et al. (2011) proposed that the term "measurement equivalence" should refer to the second set of questions, in which multi-group measurement models are compared across two or more groups. Dunbar et al. (2011) indicated that there was a lack of similarly comprehensive taxonomy and generally accepted terminology with regards to the second set of questions. The terms scalar and metric equivalence are generally accepted terms although they are not universally associated with the question whether two or more multi-group measurement models significantly differ in fit. Dunbar et al. (2011) consequently took a little bit more liberty in introducing the four hierarchical levels of measurement equivalence that are distinguished in Table 3.2. Table 3.2 presents the various forms of measurement equivalence and provides a definition of each form of equivalence.

Table 3.2

Degrees of measurement equivalence

Metric equivalence	**Scalar** equivalence	**Conditional probability** equivalence	**Full** equivalence
A multi-group measurement model in which the structure of the model is constrained to be the same across groups and in which the factor loading matrix (Λ^X) is constrained to be the same across groups does not fit multi-group data poorer than a multi-group measurement model in which the structure of the model is constrained to be the same across groups but all model parameters are freely estimated (i.e., the configural invariant multi-group model).	A multi-group measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^X is constrained to be the same across groups and in which the vector of regression intercepts (τ^X) is constrained to be the same across groups does not fit multi-group data poorer than a multi-group measurement model in which the structure of the model is constrained to be the same across groups but all model parameters are freely estimated.	A multi-group measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^X is constrained to be the same across groups, in which τ^X is constrained to be the same across groups and in which the measurement error variance-covariance matrix (Θ_δ) is constrained to be the same across groups does not fit multi-group data poorer than a multi-group measurement model in which the structure of the model is constrained to be the same across groups but all model parameters are freely estimated.	A multi-group measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^X is constrained to be the same across groups, in which τ^X is constrained to be the same across groups, in which Θ_δ is constrained to be the same across groups and in which the latent variable variance-covariance matrix (Φ) is constrained to be the same across groups does not fit multi-group data poorer than a multi-group measurement model in which the structure of the model is constrained to be the same across groups but all model parameters are freely estimated.

(Dunbar et al., 2011, p. 13)

Due to the lack of literature referring to the question whether measurement error variances differ significantly across groups, the term 'conditional probability equivalence' was coined by Dunbar et al. (2011). The term points to the fact that the conditional probability of exceeding a specific indicator variable score, given a specific standing on the latent variable of which X is the indicator, will only be the same for members of two groups if the regression of X on ξ coincides in terms of slope and intercept across the two groups and if the variance of the conditional X distributions are the same across groups (Dunbar et al., 2011).

Research on the various forms of measurement invariance and the various forms of measurement equivalence are evaluated in the hierarchical manner from left to right as presented in Tables 3.1 and 3.2 respectively, once configural invariance has been shown (Dunbar et al., 2011). Configural invariance, or lack of construct bias, is therefore a necessary prerequisite for testing weak, strong, strict and complete invariance. The test of equivalence in addition is only really meaningful if a finding of invariance has been obtained on the corresponding level of measurement invariance. Dunbar et al. (2011, p. 14) use the example that "it only really makes sense to evaluate metric equivalence if weak invariance has been shown." They further explained that a finding of invariance indicates that the multi-group model with a specific level of constraints imposed is permissible in the sense that it provides a satisfactory explanation of the observations made, specifically the observed covariance matrices. That does not answer the question of whether a multi-group model with fewer constraints imposed might not provide a more permissible explanation of the observed covariance matrices. A finding of equivalence means that the multi-group model with a specific level of constraints imposed, that provides a satisfactory account of the observations made, does not provide a significantly less satisfactory description of the observations made than a multi-group model without the constraints (Dunbar et al., 2011).

3.4 PARTIAL INVARIANCE AND PARTIAL EQUIVALENCE

The concept of partial measurement invariance and equivalence has been proposed as a compromise between full measurement invariance and equivalence and a complete lack of measurement invariance and equivalence. In the taxonomy of measurement invariance and equivalence proposed by Dunbar et al., (2011), a finding of a lack of invariance and/or a lack of equivalence will result in the termination of testing as subsequent tests assume invariance on the previous level. This, however, seems an unnecessarily strict interpretation of invariance and equivalence, both when viewing invariance and equivalence analysis from the perspective of measurement bias and cross-group comparative research. Research by Byrne et al., (1989) indicates that upon findings of lack of invariance or lack of equivalence, no follow-up procedures are

implemented or even suggested for pinpointing the sources of inequality. The unfortunate lack of tests to determine partial measurement invariance and equivalence has left a gap in the literature and consequently the impression that, given a non-invariant pattern of **Λ**, **τ**, **Θ**, or **ϕ**, further testing of invariance and equivalence are unwarranted (Byrne et al., 1989).

Practically, full measurement invariance and equivalence does not frequently occur (Steenkamp & Baumgartner, 1998). When a lack of invariance and/or equivalence is displayed, Cheung and Rensvold (1999) recommend several courses of action: (a) non-invariant items can be deleted; (b) partial invariance and partial equivalence can be utilised to retain them; or (c) non-invariant items can be interpreted as cross-group data in their own right. All these techniques require that the non-invariant items be identified correctly.

Weak invariance and probably even more so metric equivalence are sometimes difficult to achieve. Some researchers (e.g. Byrne et al., 1989; Marsh & Hocevar, 1985) have proposed relaxing these two conditions as a prerequisite for cross-group comparisons. It is recommended to rather rely upon partial weak invariance and partial metric equivalence. Under these models, non-invariant items are retained and their factor loadings are allowed to vary when analysing between-group differences (Cheung & Rensvold, 1999). If the non-invariant items constitute only a small portion of the model, then it is assumed that they will not significantly affect cross-group comparisons (Cheung & Rensvold, 1999).

In testing for measurement invariance and equivalence, if the list of items does not collectively display invariance and/or equivalence across groups, the question arises as to which items are non-invariant (Cheung & Rensvold, 1999)? The appropriateness of partial invariance and partial equivalence is determined by whether non-invariant items can be accurately identified, and the extent of their departure from invariance (Cheung & Rensvold, 1999).

Byrne et al.'s (1989) partial measurement invariance and equivalence procedure applies to factors that are configural invariant and therefore the non-invariant problem first emerges when weak invariance is imposed on the model. If the criterion for metric equivalence is not met (i.e. the weak invariance model fits significantly poorer than the less constrained configural invariance model), then additional tests are required to determine the sources of non-invariance (Cheung & Rensvold, 1999).

For a measurement model which consists of multiple constructs, each of the constructs in the model needs to be examined for invariance (Byrne et al., 1989; Cheung & Rensvold, 1999). A separate model is estimated for each construct in which the factor loadings associated with the construct are

constrained to be equal across groups while the loadings associated with the other constructs are not (Cheung & Rensvold, 1999). The statistical and practical significance of these estimated models for each construct can then be calculated in order to determine if the fit is poorer than the configural invariance model. If the change in the indices is significant, then at least one of the items within the constructs is non-invariant (Cheung & Rensvold, 1999). All non-invariant constructs are noted, and their items are examined for invariance.

Following the identification of the non-invariant constructs, Byrne et al., (1989) makes a cross-group comparison of each of the factor loadings associated with each of the non-invariant constructs. A series of tests are performed where a separate model is estimated for each item, in which the items' factor loading is constrained to be equal across groups (Cheung & Rensvold, 1999). Once again, each of the constrained models are compared to the less constrained configural invariance model, via the calculation of the statistical and practical significance. If the change in the indices is significant, then that item is non-invariant. This procedure, however, can become quite cumbersome in large measurement models.

Other procedures have been noted that can be used to identify non-invariant items. One such procedure involves examining the factor loadings of the configural invariance (unconstrained) model, and those having the greatest difference between groups are identified as non-invariant (Cheung & Rensvold, 1999). It is suggested that a limitation of this procedure involves its exclusion of significance tests of the observed differences. Alternatively, the significance of factor loadings are examined in which a loading is identified as non-invariant if it is significant for one group but not for another. Problems arise where the significance levels are nearly equal, yet one is significant and the other is not. Another procedure to decide on which loadings are invariant involves examining modification indices (MI) and expected parameter changes (EPC; Cheung & Rensvold, 1999; Steenkamp & Baumgartner, 1998). A large MI and EPC in a fully constrained model indicate that the constraint ought to be relaxed in order to improve the fit, and the item is therefore taken to be non-invariant. However, using MI and EPC would allow cross-loadings from the items on different sub-dimensions.

Two alternative procedures include the factor-ratio test and the triangle heuristic test. The factor-ratio test is a far-reaching extension of Byrne et al.'s (1989) procedure, which systematically examines all combinations of referents (the selected item's factor loading is set equal to unity across groups) and arguments (item being tested for invariance), across all groups (Cheung & Rensvold, 1999). The "triangle heuristic" is a systematic procedure that can be used to identify an invariant set of items, as this procedure proposes that an item can only be considered invariant if it belongs to an

invariant set. An item belongs to an invariant set when tested using all other members of the set as referents (Cheung & Rensvold, 1999). However, using these procedures with scales with a large amount of items, would become quite complex.

Steenkamp and Baumgartner (1998) suggest that if partial weak invariance and equivalence is supported, partial strong invariance can be tested. The intercepts of those items that are not metric equivalent across groups should be left unconstrained across groups, while the intercepts of the other items are (initially) held invariant (Steenkamp & Baumgartner, 1998). The possibility that some items have invariant loadings but different intercepts across groups does exist. If partial scalar equivalence is not initially met, non-invariant intercepts across groups can be identified and then the constraints on these intercepts relaxed in a series of further tests. Ideally, a majority of factor loadings and intercepts will be invariant across groups because in that case the latent means are estimated more reliably (Steenkamp & Baumgartner, 1998). However, according to Hair et al., (2006) partial metric and scalar invariance is achieved when at least two items do not display non-uniform or uniform DIF. Partial measurement invariance and equivalence can also be investigated for the error variance, factor variances and factor covariances. If measurement instruments are at least partially strong invariant and scalar equivalent, valid cross-group comparisons can be conducted even when the ideal of full equivalence is not realised (Steenkamp & Baumgartner, 1998). However, the question arises as to what extent is a measurement instrument sufficiently partially invariant and/or equivalent to make meaningful cross-group comparisons? Where does one draw the line in terms of relaxing non-invariant $\boldsymbol{\Lambda}$, $\boldsymbol{\tau}$, $\boldsymbol{\Theta}$, or $\boldsymbol{\phi}$? This question is especially relevant when the instrument is not used to make cross-group comparisons in structural relations or in latent means but to make cross-group comparisons in terms of observed scores. In the former case, factor loadings and/or intercepts that differ across groups can be allowed to vary across groups when specifying the group-specific structural models to compare the structural relations across groups or to compare the latent means. However, when using the instrument in a heterogeneous group to derive observed scores, items that differ in the manner in which they are related to the underlying latent variable they are meant to reflect (in terms of slope, intercept and/or error variance) need to be excluded from the calculation of the observed score. In both cases the ideal would be if the SEM software would be able to derive latent score estimates from the measurement model parameter estimates for the appropriate multi-group measurement model. In LISREL it is currently only possible to estimate latent scores for single-group measurement models but not multi-group measurement models and therefore also not partially invariant multi-group measurement models. Whether this is due to software limitations or inherent mathematical obstacles is not clear. Neither is it known whether other SEM software packages suffer from the same limitation. If it eventually would

become possible to estimate latent scores from partially invariant multi-group measurement models the problems imposed by measurement bias on cross-cultural psychometric assessment practices would thereby have been effectively and elegantly solved.

3.5 RESEARCH QUESTIONS

Measurement invariance and equivalence hypothesis testing, in the purest sense of CFA, has not been investigated in South Africa for the SWSI. By using a multi-group CFA SEM approach, the research aims to answer the following central research question:

Does the multi-group measurement model, implied by the design intentions of the developers of the SWSI, fit data obtained for a South African sample of men and women on the instrument, and are the measurement model parameters invariant and equivalent across gender subsamples?

In order to answer this broad question, the following specific research questions should be answered in the order presented:

1. When fitting the occupational stress single-group measurement model to each gender sample independently, does the single-group measurement model fit the data adequately?
2. When fitting the multi-group occupational stress measurement model to the separate gender samples, does the multi-group measurement model fit adequately when constraining the structure of the model to be equal, while all measurement model parameter estimates are allowed to vary between groups (configural invariance)?
3. When fitting the multi-group occupational stress measurement model to the separate gender samples simultaneously, does the model fit adequately when constraining the structure of the model to be equal and constraining the slope of the regression of the indicator variables on the latent variables to be equal, while all other measurement model parameter estimates are allowed to vary between groups (weak invariance)?
4. When fitting the multi-group occupational stress measurement model to the separate gender samples simultaneously with the structure of the model constrained to be equal and in which all other model parameters estimated freely but for the slopes of the regression of the indicator variables on the latent variables, does the fit of the model deteriorate significantly[14] in comparison to the fit obtained when only the structure of the model is constrained to be equal but all parameters are estimated freely (metric equivalence)?

[14] The term significantly is purposefully not described more specifically to allow the difference in fit to be assessed in terms of statistical significance and practical significance.

5. When fitting the multi-group occupational stress measurement model to the separate gender samples simultaneously, does the model fit adequately when constraining the structure of the model to be equal while all other measurement model parameter estimates are allowed to vary between groups, but for the factor loadings and the vector of regression intercepts (strong invariance)?
6. When fitting the multi-group occupational stress measurement model to the separate gender samples simultaneously, with the structure of the model constrained to be equal and in which all other model parameters are estimated freely but for the slope and the intercepts of the regression of the indicator variables on the latent variables, does the fit of the model deteriorate significantly in comparison to the fit obtained when only the structure of the model is constrained to be equal but all parameters are estimated freely (scalar equivalence)?
7. When fitting the multi-group occupational stress measurement model to the separate gender samples simultaneously, does the model fit adequately when constraining the structure of the model to be equal while all other measurement model parameter estimates are allowed to vary between groups, but for the factor loadings, the vector of regression intercepts and the measurement error variances of the indicator variables (strict invariance)?
8. When fitting the multi-group occupational stress measurement model to the separate gender samples simultaneously, with the structure of the model constrained to be equal and in which all other model parameters are estimated freely but for the factor loadings, regression intercepts and the measurement error variances of the indicator variables, does the fit of the model deteriorate significantly ($p < 0.05$) in comparison to the fit obtained when only the structure of the model is constrained to be equal but all parameters are estimated freely (conditional probability equivalence)?
9. When fitting the multi-group occupational stress measurement model to the separate gender samples simultaneously, does the model fit adequately when constraining the structure of the model to be equal and in which all other measurement model parameter estimates are constrained to be the same across the samples (complete invariance)?
10. When fitting the multi-group occupational stress measurement model to the separate gender samples simultaneously, with the structure of the model constrained to be equal and all other model parameters constrained to be equal across the samples, does the fit of the model deteriorate significantly in comparison to the fit obtained when only the structure of

the model is constrained to be equal but all parameters are estimated freely (full equivalence)?

The relevance of questions 2 to 10 is dependent on the answer obtained to question 1. If the single-group SWSI measurement model does not fit the separate gender groups, then it really makes no sense to ask the question whether the instrument measures the same construct, and whether it measures the same construct in the same way over the different groups. If the SWSI measurement models independently fit the separate gender samples it would be necessary to formally confirm the same number of factors required to satisfactorily account for the observed covariance matrix and that the nature of the loading pattern are the same across gender groups (question 2). If, but only if, configural invariance would be shown, a legitimate question to ask would be whether the structure and the factor loadings are the same across groups (question 3). If weak invariance is found, it means that it is tenable that the slope of the regression of the items on the latent variables are the same across gender groups. It does not, however, mean that a position that one or more of the slope parameters differ across groups might not be a more tenable position. A finding of weak invariance allows for the testing of a situation where the multi-group SWSI measurement model in which the structure and factor loadings are constrained to be equal, is compared to the SWSI measurement model in which only the structure is constrained to be equal (question 4). If metric equivalence is indicated, question 5 is aimed at establishing whether it would be permissible (under the more lenient interpretation of item bias) to interpret equal observed scores of individuals from different groups as indicating an equal standing on the latent variable. This would only be permissible if the regression of items on latent variables would coincide in terms of slope and intercept. Upon confirming strong invariance, comparing the measurement model in which the structure, slopes and intercepts are constrained to be equal to the measurement model in which only the structure is constrained to be equal, in order to determine whether the strong invariance multi-group model fits significantly poorer, is required (question 6). If scalar equivalence is found, it then makes sense to pose the question whether the error variances are also equal across groups (question 7). A lack of strict invariance would indicate the presence of item bias under the more stringent definition of item bias. If strict invariance is found, the measurement model in which the structure, slopes, intercepts and error variances are constrained to be equal is compared to the measurement model in which only the structure is constrained to be equal, in order to determine whether the strict invariance multi-group model fits significantly poorer and ultimately whether conditional probability equivalence exists (question 8). If conditional probability equivalence is confirmed, then it would be permissible (under the more stringent interpretation of item bias) to interpret equal observed scores of individuals from different groups as indicating an equal standing

on the latent variable. From a measurement bias perspective, it is not really necessary and relevant to test for complete invariance. If the sample comes from the same population and not from different populations complete invariance has relevance from the perspective of cross-validation. Question 9 determines whether all measurement model parameters are equal across the two gender groups. If complete invariance is confirmed, then it is appropriate to test for full equivalence which involves the comparison of the measurement model in which all parameters are constrained to be equal to the measurement model in which only the structure is constrained to be equal, in order to determine which model fits better (question 10).

Both issues of measurement invariance and measurement equivalence need to be examined if a conclusive verdict is to be made on whether the SWSI is biased or not. The next section aims to describe how these research questions will be operationalised and practically answered.

CHAPTER 4
RESEARCH METHODOLOGY AND PRELIMINARY DATA ANALYSIS

The fundamental hypotheses that were tested in this investigation is that the SWSI measures the stress construct as constitutively defined in both genders, and that the construct is measured in the same manner across gender groups. A series of confirmatory factor analyses (CFAs), in which the fit of the implied multi-group measurement model was evaluated, was conducted in order to determine the validity of the hypotheses. The validity and credibility of the verdict on the validity of these claims depends on the methodology used to arrive at the verdict. The methodology serves the epistemic ideal of science (Babbie & Mouton, 2001). If the methodology would be compromised the chances of arriving at a valid conclusion on the measurement invariance and equivalence of the SWSI would be jeopardized. The credibility of the verdict on the appropriateness of using the SWSI across gender groups included in this study, would thereby suffer. Research methodology serves the epistemic ideal through objectivity and rationality as two characteristics of the scientific method (Babbie & Mouton, 2001). Objectivity refers to a purposeful, explicit attempt to minimise error. The scientific method demands that a number of critical junctures where the risk is higher that the epistemic ideal might derail should be closely inspected, and that appropriate steps should be taken at these points to maximize the likelihood of valid findings. Science is rational in that it insists that subject matter experts should critically evaluate the validity of research findings by evaluating the methodological rigour of the process that was used to arrive at the conclusions (Babbie & Mouton, 2001). Scientific rationality is, however, contingent on an accurate description and a thorough motivation of the methodological choices that were made at the various critical junctures in the method where the epistemic ideal threatens to derail. A comprehensive description of the research methodology allows knowledgeable peers to identify methodological flaws and to point out the implication of these for the validity of the conclusions.

4.1 RESEARCH HYPOTHESES

The substantive hypotheses tested in this study was that the SWSI provides a valid and reliable measure of the occupational stress construct as defined by the instrument in both genders, and that the construct is measured in the same manner across gender. This translates to the following specific operational hypotheses:

• Operational hypothesis 1: A single-group occupational stress measurement model implied by the scoring key of the SWSI can closely reproduce the covariances observed between the individual items comprising each of the scales in the separate gender groups.

• Operational hypothesis 2: A multi-group occupational stress measurement model implied by the scoring key of the SWSI, with only the structure constrained to be equal across genders but all other measurement model parameters estimated freely, can closely reproduce the covariances observed between the individual items comprising each of the scales in the combined sample (i.e., the multi-group occupational stress measurement model implied by the scoring key of the SWSI displays configural invariance).

• Operational hypotheses 3-6: The multi-group occupational stress measurement model implied by the scoring key of the SWSI displays weak invariance, strong invariance, strict invariance and complete invariance across gender groups.

• Operational hypotheses 7-10: The multi-group occupational stress measurement model implied by the scoring key of the SWSI displays metric equivalence, scalar equivalence, conditional probability equivalence and full equivalence across gender groups.

4.2 RESEARCH DESIGN

The hypotheses formulated above make specific claims with regards to the SWSI measurement model. The occupational stress measurement model implied by the scoring key of the SWSI hypothesizes specific measurement relations between the items comprising the instrument and the latent stress dimensions measured by the instrument. More precisely, the single-group SWSI measurement model assumes that the slope of the regression of the specific indicator variables (X) on the specific latent variable (ξ), the indicator variable it is meant to represent, is positive and significantly greater than zero. Additionally, the SWSI measurement model makes assumptions about the covariance between the latent variables and the covariance between the measurement error terms. The multi-group SWSI measurement model, moreover, assumes that the intercept, slope and error variance of the regression of the specific indicator variables (X) on the specific latent variable (ξ) are the same across genders.

To empirically test the merit of the assumptions made by the single- and multi-group SWSI measurement model a plan or strategy that will direct the gathering of empirical evidence to test the operational hypotheses was required. This plan or strategy is represented by the research design

(Kerlinger & Lee, 2000), which attempts to ensure empirical evidence that can be interpreted unambiguously for or against the operational hypotheses.

This study used an *ex post facto* correlational research design. In terms of the logic of the *ex post facto* correlational design, the researcher observes the observed variables and calculates the covariance between the observed variables (Kerlinger & Lee, 2000). The observed variables could be individual items or item parcels as linear composites of individual items. Estimates for the freed single- or multi-group measurement model parameters are obtained in an iterative fashion with the purpose of reproducing the observed covariance matrix as accurately as possible (Diamantopoulos & Siguaw, 2000). If the fitted single- or multi-group model fails to accurately reproduce the observed covariance matrix or matrices (Byrne, 1989; Kelloway, 1998), the conclusion would inevitably follow that the measurement model underlying the SWSI does not provide an acceptable explanation for the observed covariance matrix/matrices. This would provide an indication that the SWSI does not measure the occupational stress domain, as intended by the measure, over the South African samples included in the study. The contrary, however, is not true. If the covariance matrix/matrices derived from the estimated model parameters closely corresponds to the observed covariance matrix/matrices it would not imply that the processes postulated by the single-/multi-group measurement model necessarily produced the observed covariance matrix/matrices, and that the SWSI therefore measures the occupational stress domain as intended. A high degree of fit between the observed and estimated covariance matrices would only imply that the processes portrayed in the measurement model provide one plausible explanation for the observed covariance matrix/matrices.

4.3 STATISTICAL HYPOTHESES

The nature of the statistical analyses that was used to test the operational hypotheses affected the decision as to whether statistical hypotheses should be formulated, and the format in which they were formulated. One possibility would have been to use an unrestricted, exploratory factor analytic approach in which no *a priori* stance is taken on the number of factors underlying the observed covariance matrix, nor on their identity and the manner in which the items load on the factors (Ferrando & Lorenzo-Seva, 2000). If this option would have been chosen, no statistical hypotheses would have been formulated. This option ignores the design intentions of the developers of the SWSI, and therefore seemed inappropriate.

In the case of the SWSI, a very specific stance is taken on the number of stress factors underlying the observed covariance matrix, their identity and the manner in which the items load on the stress

factors. Occupational stress items were explicitly and intentionally developed to reflect specific dimensions of the occupational stress construct. Specific SWSI items were written to function as stimulus sets to which test takers would respond with behaviour which would be behavioural expressions of specific latent occupational stress dimensions. The scoring key of the SWSI reflects these design intentions.

Therefore, it seemed more reasonable towards the developers of the instrument to first evaluate the question whether the intentional instrument design succeeded in providing a comprehensive and relatively uncontaminated empirical grasp on the occupational stress construct as the SWSI manual defines it. A hypothesis-testing, restricted, confirmatory factor analytic approach should rather be followed. In terms of this approach specific structural assumptions are made with regards to the number of latent variables underlying the SWSI, the relations among the latent variables and the specific pattern of loadings of indicator variables on these latent variables (Ferrando & Lorenzo-Seva, 2000; Jöreskog & Sörbom, 1993). Specific assumptions are, moreover, made on how these structural assumptions apply across gender groups.

Structural equation modelling utilising LISREL 9.00 (Jöreskog & Sörbom, 1996a) was used to test the ten operational hypotheses listed in paragraph 4.1. Six of these eight hypotheses were translated into statistical hypotheses.

The exact fit null hypothesis was not tested. The exact fit null hypothesis represents the somewhat unrealistic position that the single-group measurement model is able to reproduce the observed covariance matrix to a degree of accuracy that could be explained in terms of sampling error only. Browne and Cudeck (1993, p. 137) consequently argued that “in applications of the analysis of co-variance structures in the social sciences it is implausible that any model that we use is anything more than an approximation to reality. Since a null hypothesis that a model fits exactly in some population is known *a priori* to be false, it seems pointless even to try to test whether it is true”. Operational hypotheses 1 and 2 in addition explicitly assume that the measurement model only approximates the processes that operated in reality to create the observed covariance matrix (Browne & Cudeck, 1993):

Operational hypothesis 1 was tested by testing the following close fit null hypotheses

$$H_{01a\_male}: \text{RMSEA} \leq .05$$

$$H_{a1a\_male}: \text{RMSEA} > .05$$

H_{01b_female}: RMSEA ≤ .05

H_{a1b_female}: RMSEA > .05

Operational hypothesis 2 was tested by testing the following close fit null hypothesis:

H_{02}: RMSEA ≤ .05

H_{a2}: RMSEA > .05

Operational hypothesis 3-6 was tested by testing the following close fit null hypotheses:

H_{0i}: RMSEA ≤ .05; i=3, 4, 5, 6

H_{ai}: RMSEA > .05; i=3, 4, 5, 6

No formal statistical hypotheses were formulated for operational hypotheses 7-10. Operational hypotheses 7-10 were evaluated via the practical significance of the difference in fit between the configural invariance model and the weak, strong, strict and complete invariance models respectively in terms of the three fit indices proposed by Cheung and Rensvold (2002).

H_{0i}; i=1, 2, .., 6 were rejected if the conditional probability associated with the sample RMSEA estimate under the close fit null hypothesis was equal to or less than .05. If the conditional probability associated with the sample RMSEA estimate under the close fit null hypothesis was greater than .05, H_{0i} would not be rejected.

Operational hypotheses 7-10 would be rejected (i.e., the difference in fit between the configural invariance model and the weak, strong, strict and complete invariance models will be considered practically significant) if a change of more than -.01 in the CFI fit index, a change of more than -.001 in the Gamma Hat fit index (Γ_1) and a change of more than -.02 in the McDonald Non-centrality index (Cheung & Rensvold, 2002) was observed between the partially constrained multi-group model and the fully unconstrained multi-group model. If ΔCFI, ΔGamma Hat and ΔMc was less than or equal[15] to the critical thresholds for any of the model comparisons, equivalence would be assumed confirmed.

If H_{01a_male} and H_{01b_female} would not be rejected, indicating close single-group model fit, then a further series of hypotheses on the slope of the regression for the individual items on the respective latent stress dimensions would be tested.

[15] Less than or equal to the critical thresholds means values closer to zero.

The following 59 null hypotheses on the slope of the regression of individual item j on latent work stress dimension k were tested:

$$H_{oi}: \lambda_{jk}=0;\ i=7, 8, \ldots, 65;\ j=1, 2, \ldots, 59;\ k=1, 2, \ldots, 9$$

$$H_{ai}: \lambda_{jk}\neq 0;\ i=7, 8, \ldots, 65;\ j=1, 2, \ldots, 59;\ k=1, 2, \ldots, 9$$

The results of these analyses formed the basis for examining the merits of the claim made by the developers of the test that the SWSI via the occupational stress scales (i.e. the General Work Stress scale and the eight Sources of Work Stress scales) successfully measures the work stress construct it intends to measure, and in the manner that it intends to do so according to the scoring key.

4.4 SAMPLE

Determination of the sample size necessary to achieve adequate power is an important issue. Sample sizes of at least 200 observations are regarded as satisfactory for most SEM applications (Kelloway, 1998). For a study that intends using SEM, three issues should be considered when deciding on the appropriate sample size.

The first consideration was the ratio of sample size to the number of parameters to be estimated. A situation in which more freed model parameters have to be estimated than there are observations in the sample would be regarded as unacceptable. Larger sample sizes are required for more elaborate measurement models which contain more variables and therefore have more freed parameters that have to be estimated. Likewise, larger samples are required to fit multi-group measurement models, and especially the multi-group configural invariance model. Bentler and Chou (as cited in Kelloway, 1998, p. 20) recommend that the ratio of sample size to number of parameters estimated should fall between 5:1 and 10:1. The SWSI single group measurement model and multi-group configural invariance measurement model would, in terms of the Bentler and Chou (as cited in Kelloway, 1998) guideline, require a sample of 1065-2130 research participants and 2130-4260 research participants respectively, to provide a convincing test of the measurement model (213 freed parameters and 426 freed parameters respectively).

The second consideration to take into account when deciding on the appropriate sample size involved the statistical power associated with the test of the hypothesis of close fit (H_{0i}: RMSEA ≤ .05; i=1, 2, .., 6) against the alternative hypothesis of mediocre fit (H_a: RMSEA > .05; i=1, 2, .., 6). In the context of SEM, statistical power refers to the probability of rejecting the null hypothesis of close fit (H_{0i}: RMSEA ≤ .05; i=1, 2, .., 6) when in fact it should not be rejected (i.e., the model fit actually is mediocre, H_{ai}: RMSEA > .05; i=1, 2, .., 6). Exceedingly high statistical power would mean that any

attempt to formally empirically corroborate the validity of the model would be futile. Even a small deviation from close fit would result in a rejection of the close fit null hypothesis. Conversely, however, exceedingly low power would mean that even if the model fails to fit closely, the close fit null hypothesis would still not be rejected. Not rejecting the close fit null hypothesis under conditions of low power will therefore not provide very convincing evidence supporting the validity of the model. Power tables, compiled by MacCallum, Browne and Sugawara (1996), were used to derive sample size estimates for the test of close fit, given the effect sizes assumed above, a significance level (α) of .05, a power level of .80 and degrees of freedom (ν) of (½[(p+q][p+q+1]-t)=1829-213=1616. The MacCallum et al. (1996) table indicated that a minimum sample of 132 observations would be required to ensure statistical power of .80 in testing the null hypothesis of close fit for the SWSI single-group measurement model. The MacCallum et al. (1996) power table, however, only makes provision for models with degrees of freedom up to 100. Syntax developed by Preacher and Coffman (2006) in R (and available at http://www.quantpsy.org/rmsea/rmsea.htm) was also used to determine the sample size required to ensure a statistical power of .80 for the test of close fit. For this purpose, a significance level of .05 was specified, 1616 degrees of freedom, RMSEA was set to .05 under H_0 and RMSEA was set to .08 under H_a. The Preacher and Coffman (2006) software returned a required sample size of 25.68359 cases.

The third consideration to take into account when deciding on the appropriate sample size is practical and logistical considerations like cost, availability of suitable respondents and the willingness of a test distributer company to commit data from a large archival database.

For this study, the archival data have been provided by a test distributor company in an anonymous format[16]. The purpose of SWSI assessment in most cases would have been for research projects and company wellness interventions (N. Taylor, personal communication, 4 September 2012). Approximately 60% of the cases contained both the age and race of respondents, and approximately 70% of the records included the race of respondents, however the majority of the cases did not include all biographical information (i.e., race, age and gender) together. Biographical information with regards to education level, current occupation or first language was not included in the data. The lack of biographical information was rather unfortunate as it prevented the proper characterisation of the study sample. This was certainly a regrettable shortcoming in this study. As this research aimed to determine the invariance and equivalence of the measurement model of the

[16] Written institutional permission has been obtained from the test distributor to utilise the data of the sample for the purpose of this research.

SWSI across gender, the sample was considered suitable for the proposed purpose as gender information was provided by the questionnaire distributor for all the selected cases.

The SWSI sample consisted of 1209 respondents where gender was indicated. Of this sample, 749 (62%) were male and 460 (38%) were female. Given that, some goodness-of-fit measures (e.g., the chi-square statistic) are known to be affected by sample size (Cheung & Rensvold, 2002; Diamantopoulos & Siguaw, 2000; Marsh, H.W., Hau, K., Balla, J.R., & Grayson, D., 1998), the sample sizes for the male and female groups were matched. Hence, 460 male cases were randomly selected using SPSS from the original male sample. It would have been preferable to have matched the samples on the mean age, however the lack of biographical information did not allow for this as the sample size would have decreased substantially. Exploring the complexities regarding the subsets of each male/female sample would be difficult as there was a significant amount of missing information for selected cases and therefore reporting other sample characteristics would be futile. However, the gender details were the primary concern in this study and will remain the focus of how the data was handled.

4.5 STATISTICAL ANALYSES

Structural Equation Modelling (SEM) was used to perform a series of confirmatory factor analysis on the subscales of the SWSI using LISREL 9.00 (Du Toit & Du Toit, 2001; Jöreskog & Sörbom, 1996a). SEM is a collection of statistical techniques that examine the relationship between one or more independent and one or more dependent latent variables and/or between one or more independent latent variables and one or more dependent observed variables (Davidson, 2000). The variables can be examined continuously or discretely (Davidson, 2000). Kelloway (1998) argued in support of SEM, stating that (a) SEM allows the researcher to determine how well these measures reflect the intended constructs, (b) SEM permits the testing and specification of more complex path models in addition to testing the components comprising the model to make thorough predictions, and (c) it provides a flexible, yet powerful method that caters for the quality of measurement, which is very important in the evaluation of the predictive relationships amongst the underlying latent variables. Due to the above mentioned argument, this study selected SEM as a statistical analysis technique.

4.5.1 PREPARATORY PROCEDURES

The following section aims to describe and motivate the initial procedures that were undertaken prior to conducting the SEM analyses. The section begins by specifying the respective models that were subjected to confirmatory factor analyses. Thereafter, the identification of the measurement models is evaluated, and the approach used in handling missing values is indicated. The necessity of

performing item and dimensionality analyses is explained and the procedures described. This is followed by a discussion on fitting the measurement model as well as the assessment of discriminant validity. Finally, the procedures for investigating measurement invariance and equivalence are discussed and explained in section 4.5.2.5.

4.5.1.1 MODEL SPECIFICATION

The detailed specification of the measurement models, in SEM notation, was required to determine whether the relevant measurement models are identified. The specification provides a clear understanding of the model complexity as well as the number of parameters to be estimated.

The basic first order single-group factor analysis model specification is given by Equation 1:

$$\mathbf{X} = \tau + \Lambda^{X}\xi + \delta \quad \text{----------(1)}$$

Where:

- **X** is the column vector of observable indicator scores;
- τ is a vector of intercept terms;
- Λ^{X} is the matrix of factor loadings;
- ξ is the column vector of latent factors;
- **δ** is the column vector of unique/measurement errors components comprising the combined effect on X of systematic non-relevant influences and random measurement error (Jöreskog & Sörbom, 1996a).

The foregoing measurement model implies two additional matrices. The first is the symmetrical variance-covariance Φ matrix. This matrix describes variance in, and covariance/correlations between, the latent variables in the model. All the elements of Φ are freed to be estimated. The second matrix is the diagonal variance-covariance matrix Θ_{δ} which would imply that the measurement error terms are assumed to be uncorrelated across the indicator variables. By freeing off-diagonals in this matrix, it would then imply that that error terms may be correlated, indicating the possibility of additional common factors. Due to the confirmatory nature of this study, freeing the off-diagonals would be impossible to justify in terms of the design intentions of the developers of the instrument.

The basic first order multi-group factor analysis model specification is given by Equation 2:

$$\mathbf{X}^{g} = \tau^{g} + \Lambda^{Xg}\xi^{g} + \delta^{g} \quad \text{----------(2)}$$

Where:

- $\mathbf{X}^g$ is the column vector of observable indicator scores for the gth group; g=1, 2;
- τ^g is a vector of intercept terms for the gth group; g=1, 2;
- Λ^{xg} is the matrix of factor loadings for the gth group; g=1, 2;
- ξ^g is the column vector of latent factors for the gth group; g=1, 2;
- $\boldsymbol{\delta}^g$ is the column vector of unique/measurement errors components comprising the combined effect on X of systematic non-relevant influences and random measurement error for the gth group; g=1, 2 (Jöreskog & Sörbom, 1996a).

The variance-covariance matrices Φ^g remain symmetrical matrices with all elements freed to be estimated. The variance-covariance matrices $\Theta_\delta{}^g$ remain diagonal matrices.

4.5.1.2 MODEL IDENTIFICATION

In evaluating the identification of the model, the researcher determined whether sufficient information would be available to obtain a unique solution for the parameters to be estimated in the measurement model (Diamantopoulos & Siguaw, 2000).

Diamantopoulos and Siguaw (2000) and MacCallum (1995) make two recommendations regarding model identification. Firstly, it is recommended that each latent variable should be allocated a definite scale. Secondly, the model parameters to be estimated may not exceed the number of unique variance/covariance terms in the sample observed covariance matrix. The model should therefore have positive degrees of freedom.

In the case of this study for the single-group analysis there were [(59 x 60)/2] + 59 = 1829 unique variance and covariance terms in the inter-item covariance matrix. For the multi-group analysis there was 3658 unique variance and covariance terms. The degrees of freedom for the single-group measurement model was therefore 1829-213 = 1616. The degrees of freedom for the multi-group measurement invariance models are shown in Table 4.1. Both the abovementioned requirements for model identification were therefore adhered to.

Table 4.1

Degrees of freedom for the multi-group measurement invariance models

Configural invariance	Weak invariance	Strong invariance	Strict invariance	Complete invariance
3658-426 = 3232	3658-376 = 3282	3658-317 = 3341	3658-258 = 3400	3658-213 = 3445

4.5.1.3 TREATMENT OF MISSING VALUES

Missing values needed to be identified and handled to ensure the completeness of the data prior to conducting the analyses. The missing values analysis was conducted using the PRELIS software. Missing values can be handled in numerous ways, including: (a) listwise deletion, (b) pairwise deletion, (c) mean substitution, (d) group mean substitution, (e) imputation by regression, (f) imputation by matching, (g) expectation maximisation, (h) full information maximum likelihood and (i) multiple imputation (Du Toit & Du Toit, 2001).

The options of listwise deletion, pairwise deletion, mean substitution, group mean substitution and imputation by regression were not seriously considered as possible solutions to the missing values problem in this study due to the availability of more sophisticated procedures.

Imputation by matching is an approach that replaces a missing value with an actual value from one or more similar cases in the current dataset (Kline, 2005; Olinsky, Chen & Harlow, 2003). The technique separates complete from incomplete cases, then sorts both sets of records so that cases with similar profiles on matching variables (determined by the researcher) are grouped together. The incomplete record is then randomly included among the complete records, and missing scores are replaced with those in the same variable from the nearest complete record. This process continues until the case contains no missing data (Kline, 2005). In the case of imputation by matching (Jöreskog & Sörbom, 1996a), the imputation of a missing value on variable y_a for a specific case *a* with no missing values on a set of p matching variables x_1, x_2 , …, x_p involves the following procedure:

- All cases b_i; i=1, 2, …, n are identified with no missing values on either y_{bi} or on the set of matching variables for which $W = \Sigma(zb_i - za_i)^2$; i=1, 2, …, n is a minimum.
- If only n=1 case exists for which W is a minimum, then y_a is simply replaced by y_b.
- If, however W is a minimum for n > 1 cases, with y values $y_1^{(m)}$, $y_2^{(m)}$, …,$yn^{(m)}$, the mean $E(y^m)=(1/n)\Sigma y_i^{(m)}$ and variance $s^2_m =(1/[n-1])\Sigma(y_1^{(m)}-E(y^m))$ of the y-values of the matching cases will be calculated.
- If $s^2_m/s^2_y < v$, where the variance ratio *v* was set equal to 0.50, y_a is replaced by $E(y^m)$. If the variance ratio does not pass the critical value, no imputation is done (Jöreskog & Sörbom, 1996a).

Supporters of the imputation by matching approach indicate that the imputed values preserve the distributional characteristics of the data as opposed to a mean substitution. A pitfall to this approach

is that when dealing with large datasets, many matching variables would need to be specified, making the decisions regarding matching cumbersome. It is also ideal that the matching variables should not be variables that will be included in the actual data analysis (Dunbar-Isaacson, 2006; Olinsky et al., 2003).

With multiple imputation the iterative estimation process is replicated between five to ten times (Olinsky, A., Chen, S., & Harlow, L. (2003). Each replication produces respective datasets of imputed values. Each dataset is then used to estimate the measurement/structural model. Due to the variability in the datasets, it is then possible to estimate standard errors. The advantage of the multiple imputation procedure available in LISREL 9.00 is that estimates of missing values are derived for all the cases in the initial sample (i.e., no cases with missing values are deleted), and the data set is available for subsequent item and dimensionality analyses. The multiple imputation procedure assumes that the data is missing at random (MAR). Data is missing at random when the probability of missing data on any variable is not related to its particular value but the pattern of missing data is predictable from other variables in the database.

Multiple imputation and full information maximum likelihood (FIML) are more elegant and mathematically sophisticated procedures for replacing missing values. However, these methods make more stringent assumptions with regards to the data, which include that the variables are continuous and the data follows a multivariate normal distribution. According to Mels (2007) it would be acceptable to use multiple imputation and full information maximum likelihood (FIML) if observed variables are measured on a scale comprising 5 or more scale values, if the observed variables are not excessively skewed (even though the null hypothesis of multivariate normality might have been rejected), and if less than 30% of the data constitutes missing values. Although the FIML estimation procedure is more efficient than the available multiple imputation procedures in PRELIS, it has the disadvantage that no separate imputed data set is created, which thus prevents the needed preliminary analyses on the imputed data.

Based on the above discussion it was concluded that multiple imputation or imputation by matching would be the most suitable approaches for this study. The procedure chosen, however, depended on the nature of the data. Multiple imputation was therefore used to treat the missing values problem in this study. The choice of this procedure is motivated in section 5.2.

4.5.1.4 ITEM ANALYSIS

The objective of item analysis is to gain a more powerful understanding of tests or questionnaires (Murphy & Davidshofer, 2005). The procedure is essentially an analysis of correlations between each

item with a total score (Kline, 1994) as well as inter-item correlations (Murphy & Davidshofer, 2005). Test developers are likely to construct tests that would generally aim to have items that correlate on a specific scale of investigation. Items with higher correlations are assumed to be measuring the same latent variable. When developing tests/questionnaires, Nunnally (1978) indicates that item analysis is to be used to make the first item selection, and then the selected items are to be subjected to factor analysis.

For this study, item analysis was conducted as a valuable precursor to fitting the *a priori* measurement model to the data. The item analysis helped to identify whether the observed variables were consistent measures of the intended latent variable. High reliability of the measures of the intended latent variable would give credibility to the design intentions of the test developers. While Nunnally (1978) indicates that item analysis assists in making final item selection decisions, the intention of this study was to retain all items, but report on those that may be possible culprits that contribute to poor latent variable representation and possible poor model fit. Furthermore, the analyses also provided initial information regarding the homogeneity of each subscale.

For these analyses, each gender sample's data were analysed separately, thereby providing some initial information regarding reliability of the observed variables across genders. This procedure provided valuable information regarding the measurement properties of the instrument across genders. The SPSS 19 Scale Reliability Procedure was used to analyse the subscale items.

4.5.1.5 DIMENSIONALITY ANALYSIS

When constructing scales, the design intention is that the items selected to represent each latent variable would be in fact measuring the intended latent variable exclusively. This is termed the uni-dimensionality assumption (Hair et al., 2006). Strict uni-dimensionality will seldom, if ever, be achieved. Essentially, uni-dimensionality would be achieved if the partial inter-item correlations would become negligibly small when controlling for a single underlying factor (Hair et al., 2006). Investigating whether the number of factors required to satisfactorily explain the observed correlation matrix corresponds to the design intention underlying the scale, and investigating whether the resultant factor loadings are high, is an approach to take when testing this assumption. Scales that fail the uni-dimensionality assumption (i.e., more than one factor emerges naturally for a scale that was designed to measure a single latent variable) imply that multiple dimensions should be specified for the instrument. Again, testing this assumption does not negate the necessity of the CFA. Rather, it provides further insight into the internal function of the *a priori* specified factor structure of the SWSI and reasons for possible poor model fit.

The dimensionality analyses were conducted by subjecting each work stress scale to an unrestricted principle axis factor analysis with oblique rotation. Oblique rotation was chosen over varimax rotation as it is considered the superior method that can provide simple structure even when underlying factors may be related to each other (Kerlinger & Lee, 2000; Stewart, 2001). However, oblique rotation can be complex to interpret (Tabachnick & Fidell, 1989). The above analysis was performed on each of the scales individually for each gender. Principle axis factor analysis was chosen over principle components analysis. Principle components analysis does not separate error and specific variance (Kline, 1994) whereas principle axis analysis does allow for the presence of measurement error. Human behaviour without measurement error is unlikely (Stewart, 2001).

The possibility that artefact factors, which reflect differences in item difficulty value or some other descriptive characteristic of the subscale item data set, exist could be extracted during the above analyses when performing analyses on a matrix of product moment correlations (Hulin, Drasgow & Parsons, 1983). Descriptive statistics were consequently calculated for the items of each work stress scale.

In cases where uni-dimensionality was not met, the possibility of meaningful factor fusion was investigated. The question therefore is whether the extracted factors constitute meaningful subthemes within the original latent work stress dimension. In the case of sub-scales where the uni-dimensionality assumption was challenged, irrespective of whether meaningful factor fission occurred, the ability of a single factor to account for the observed inter-item correlation matrix was also investigated. This approach was taken to investigate the magnitude of the factor loadings when a single factor (as per the *a priori* model) was forced, and to examine the magnitude of the factor loadings.

In all cases, irrespective of whether the uni-dimensionality assumption was rejected, the credibility of the extracted factor structure as an explanation of the observed inter-item correlation matrix was evaluated by examining the matrix of residual correlations. The percentage of large residual correlations in the latter solution could be regarded as reflecting on the credibility of the extracted factor solution as an explanation for the observed correlation matrix.

SPSS 19 was used for the principal factor analyses described above. The eigenvalue-greater-than-unity rule of thumb was used to determine the number of factors to extract.

The objective of dimensionality analyses is to test whether the uni-dimensionality assumption is met for each factor. Inadequate factor loadings would suggest that items should be removed, and factor fission would suggest that a split should be proposed in the sub-scale factor composition. If these

actions are taken then a revision of the measurement models would take place. It is, however, important to note that the researcher does not have intellectual property rights on the instrument and does not have any mandate from the test developer to modify the instrument and its design intention. The mandate of this research is, therefore, not to redesign the measurement model in any way. Consequently, the dimensionality investigation is, in this case, not a step in ensuring that the individual items are internally consistent observational reflections of the latent work stress variables as proposed by the test authors. Rather the dimensionality analysis provides further insight into the internal function of the *a priori* specified factor structure of the SWSI and reasons for possible poor confirmatory factor analysis model fit.

The separate gender group sample results are presented. Differences between each gender sample are also discussed. While this does not provide information regarding the configural invariance of the SWSI, it does provide valuable information that could be returned to when wanting to identify reasons for poor model fit.

4.5.2 STRUCTURAL EQUATION MODELLING

4.5.2.1 VARIABLE TYPE

The SWSI utilises a five-point Likert-type response scale, and the respondent is requested to indicate their degree of preference, or level of self-perceived skill, based on item content. The data produced by this type of response scale should, strictly speaking, be regarded as ordinal data/discrete variables. Based on the results of a Monte Carlo study by Muthén and Kaplan (1985) it is, however, standard practice to specify the data obtained from Likert scales with five or more scale points as continuous data, for the purpose of CFA (Maximum Likelihood) SEM analyses. Another strategy to convert ordered categorical data to continuous data is to use item parcels rather than item-level raw data.

In the case of this study, the use of item parcelling was not a practical measure as the number of measurement model parameters that had to be estimated did not need to be reduced. Furthermore, disadvantages of using item parcelling do exist. For example Meade and Lautenschlager (2004) reported that the measurement invariance and equivalence tests of equality of factor loadings (i.e., metric invariance) tend to be more precise when using item level data. In a further study, Meade and Kroustalis (2006) found that the use of items versus item parcels is preferred when conducting tests of measurement invariance and equivalence. From their simulation studies it was found that even though fit could be poor when using item data, lack of equivalence may be masked by using item parcels. Sass and Smith (2006) also indicate that there seems to be a lack of evidence of one

single suitable approach to constructing parcels. Kim and Hagtvet (2003) indicate that using parcels may increase the likelihood of misrepresenting the latent construct. Therefore, due to the above disadvantages and the size of the SWSI work stress scales measurement model (59 items), it was decided that individual items would be a suitable strategy to employ in this study.

4.5.2.2 EVALUATION OF MULTIVARIATE NORMALITY

As described above, it is standard practice to specify the data obtained from Likert scales with five or more scale points as continuous data. The SWSI uses a five point Likert scale, therefore the individual indicator variables are regarded as continuous data. When using continuous data in SEM, maximum likelihood estimation is preferred. Other estimation methods include generalised least squares (GLS) and full information maximum likelihood (FIML). FIML is useful when dealing with missing values. However, with all these estimation methods multivariate normality is assumed for the data (Mels, 2003).

In the event of working with non-normal data, Mels (2003) indicates that alternative estimation methods should be utilized, for example: robust maximum likelihood (RML); weighted least squares (WLS); or diagonally weighted least squares (DWLS). These methods are advantageous as the interpretation of the solution is not based on transformed values (Du Toit & Du Toit, 2001). Mels (2003) does make a further recommendation that RML would be the preferred approach when dealing with multivariate non-normal data.

The normality of the individual indicators in this study was evaluated using PRELIS (Jöreskog & Sörbom, 1996b). In the case where the null hypothesis of multivariate normality was rejected, normalisation was attempted (Jöreskog & Sörbom, 1996b). If the hypothesis of multivariate normality was still rejected, robust maximum likelihood estimation was used (Mels, 2003).

4.5.2.3 MEASUREMENT MODEL FIT

In order to meet the measurement invariance and measurement equivalence research objectives of this study, LISREL 9.00 (Du Toit & Du Toit, 2001, Jöreskog & Sörbom, 1996a) was used to determine the fit of: (a) the single group occupational stress model on the two gender samples separately and (b) the multi-group occupational stress model when fitted in a series of multi-group analyses. The fit of the single- and multi-group measurement models were evaluated by testing H_{0i}: RMSEA $\leq .05$; i=1, 2, ..., 6.

4.5.2.4 DISCRIMINANT VALIDITY

Discriminant validity is the extent to which a construct may be considered to be truly distinct from other constructs given the manner in which the constructs are measured (Hair et al., 2006). Thus, high discriminant validity provides evidence that a construct is unique and captures some phenomena other measures do not capture (Hair et al., 2006). If a latent variable is not able to account for more variance in its associated observed variable than unmeasured influences (measurement error) and other constructs within the conceptual framework, then the validity of the individual indicator and of the construct is questionable (Farrell, 2010). Therefore, inferences made regarding relationships between constructs under investigation may be incorrect if discriminant validity is not shown. The presence of cross-loadings (suggested by the modification indices calculated for Λ^{X}) indicates a discriminant validity problem. If high cross-loadings do indeed exist, and they are not represented by the measurement model, the CFA fit should not be good.

For the purpose of assessing discriminant validity, the use of the CFA correlation matrix (Φ) is recommended, as this takes measurement error into account (Farrell, 2010). This minimises misleading results and provides a more stringent evaluation of discriminant validity. There are three recommended tests to examine the discriminant validity of scales. The first test is to compare the average variance-extracted proportions[17] for any two constructs with the square of the correlation estimate between these two constructs (shared variance; Farrell, 2010; Hair et al., 2006). The variance-extracted estimates should be greater than the squared correlation estimate, indicating that a latent construct explains its item measures better than it explains another construct, and thereby providing support for discriminant validity[18].

The second test compares the chi-square statistic to assess the discriminant validity of one scale with respect to another scale. In the unconstrained, alternative hypothesis model all elements in Φ are freely estimated. In the constrained, null hypothesis model one or more elements in Φ are constrained to be equal. A chi-square difference test is then used to test the null hypothesis that the constrained model does not fit the data statistically significantly poorer than the unconstrained model (Mels, 2010). The test statistic value for the chi-square difference test is the difference between the goodness-of-fit chi-square test statistic values of the single-group measurement models under the null and alternative hypotheses (Mels, 2010). The related degrees of freedom

[17] The average variance extracted proportion refers to the proportion of variance in the indicator variables that is due to the construct rather than to measurement error. The proportion average variance extracted is calculated as pv = $(\Sigma\lambda^2)/[\Sigma\lambda^2 + \Sigma(\theta_\delta)]$ (Diamantopoulos, A., & Siguaw, 2000, p. 91)

[18] The average variance extracted should also exceed at least .50 so that the latent variable being measured by the indicators/items account for a larger proportion of the variance in the indicators than measurement error.

indicate the difference between the degrees of freedom of the measurement models under the null and the alternative hypotheses.

The third test for discriminant validity involves calculating the 95% confidence interval estimate. This involves examining the correlation coefficient between the latent variables of interest (Φ). In other words, a 95% confidence interval estimate is used to test the null hypotheses of a perfect correlation versus the alternative hypotheses of a non-perfect correlation (Mels, 2010). If the latent variables of interest are indeed distinct, the correlation between them should be less than one.

Farrell (2010) suggests certain steps for when discriminant validity issues arise. Discriminant validity can be improved by removing items that cross-load on more than one latent variable. This can be performed via EFA or through CFA in which modification indices or correlated error terms are examined (Farrell, 2010). If discriminant validity issues continue, constructs can be combined into a single latent variable. This technique, however, may not always be appropriate. Collecting additional data may also be done in order to determine if discriminant validity issues are a result of sampling accidents. While Farrell (2010) makes certain suggestions to eliminate discriminant validity issues, the intention of this study would only be to report the findings of the analysis and to identify possible reasons for poor model fit.

Two of the three discriminant validity tests described above were employed in this study: the average variance extracted versus shared variance test; and calculating the 95% confidence interval estimate. Assurance for discriminant validity increases if discriminant validity is established in both of these recommended tests (as opposed to performing only one test). The tests were performed after step 1, which involves fitting the single-group measurement model (described in section 4.5.2.5).

For these analyses, each gender sample's data were analysed separately, thereby providing some initial information regarding the validity of the observed variables across gender. These procedures provided valuable information regarding the measurement properties of the instrument across gender and could be returned to when wanting to identify reasons for poor model fit.

4.5.2.5 TESTING FOR MEASUREMENT INVARIANCE AND EQUIVALENCE

This study used the specific measurement invariance and equivalence sequence of tests set out by Dunbar et al. (2011) to answer a sequence of research questions that examine the extent to which the multi-group measurement model may be considered measurement invariant and equivalent or not, and to determine the source of variance if it exists (Vandenberg & Lance, 2000). The following

series of steps capture the essential logic underlying the investigation of measurement invariance and measurement equivalence as set out by Dunbar et al. (2011).

Step 1: Establish if the single-group measurement model, when fitted to each sample independently, displays reasonable fit.

Prior to establishing the source of measurement invariance and equivalence it is necessary to first establish whether the single-group measurement model fits both gender groups separately. This step determines whether the single-group measurement model would display reasonable fit when fitted to each group independently (Dunbar et al., 2011). Rejecting the null hypothesis of close fit (H_{02}: RMSEA ≤ .05) would imply that the measurement model does not adequately fit the data of one or both samples, and any further examination of measurement invariance and measurement equivalence would be questionable (Dunbar et al., 2011). Satisfactory model fit for both gender samples would justify further measurement invariance and equivalence analyses. Starting the analyses by fitting the configural invariance multi-group model can result in ambiguous findings when the single-group model fails to fit the data. Lack of multi-group measurement model fit can be due to the measurement model not being applicable to one of the groups or to it not being applicable to both groups.

Step 2: Establish if the multi-group measurement model in which the structure of the model is constrained to be the same across groups, but with no freed parameters constrained to be equal across groups, displays reasonable fit when fitted to the samples simultaneously in a multi-group analysis.

This step involves the investigation of configural invariance (Dunbar et al., 2011). Configural invariance is a prerequisite for evaluating further aspects of measurement invariance and measurement equivalence. If there is a lack of configural invariance, other tests of measurement invariance and equivalence are unnecessary because it indicates that the measuring instrument represents different constructs across groups (construct bias). Finding support for configural invariance signifies that the different groups used the same conceptual frame of reference when they responded to the items. The measuring instrument therefore reflects the same underlying constructs across the groups. Thus, configural invariance focuses on the theoretical structure of the measurement instrument. The underlying theoretical structure of the instrument refers to the manner in which the subscales of the instrument tap into the same underlying constructs across groups (Theron, 2011). Configural invariance will most probably not be achieved if the constructs are very abstract and culture specific, and when different groups use different frames of references

when attaching meaning to the construct of interest, since these different frames of reference would probably result in the construct expressing itself in different behavioural denotations (Cheung & Rensvold, 2002). Other reasons why configural invariance may not be attained includes data collection problems and translation errors. The configural invariance model is used as the baseline model against which further nested models are evaluated (Vandenberg & Lance, 2000).

Step 3a: Establish whether the multi-group measurement model in which the structure of the model is constrained to be the same across groups, and in which all parameters are estimated freely across the samples, but for the slopes of the regression of the indicator variables on the latent variables that are constrained to be equal, demonstrates acceptable fit when fitted to the samples simultaneously in a multi-group analysis.

Upon (a) finding acceptable model fit for both samples independently, and (b) when configural invariance is supported, it is necessary to explore whether invariance exists in the factor loadings of the items on the latent variables across samples. Weak invariance needs to be tested. A lack of weak invariance would imply that the slope of the regression of at least some of the items on the latent variable they represent, differ across samples. This indicates that the item content is being perceived and interpreted differently across samples (Byrne & Watkins, 2003). This would be a disappointing result in measurement invariance and equivalence research as the factor loadings reflect the foundation of the measurement process (Dunbar et al., 2011). Finding support for weak invariance would be a suitable result as it would support the position that the items operate in approximately the same way across samples in the way they reflect the underlying latent variables they are meant to reflect (Dunbar et al., 2011). If weak invariance (lack of non-uniform item bias) has been established, then metric equivalence will be tested. If a lack of weak invariance (non-uniform item bias) has been established, the process will then be terminated[19] since, irrespective of what further tests might reveal, the measures are biased and observed score differences across genders cannot be interpreted to indicate corresponding differences in the latent variable.

Step 3b: Establish whether the multi-group measurement model in which the structure of the model is constrained to be the same across groups, and in which all parameters are estimated freely across the samples, but for the slopes of the regression of the indicator variables on the latent variables fits the multi-group data poorer than a multi-group measurement model in which the structure of the model is constrained to be the same across groups, but all parameters are estimated freely.

[19] The process was not in fact terminated as this study further investigated partial weak invariance in order to determine which slopes are biased. Biased slopes were therefore identified and acknowledged in further tests of partial invariance and partial equivalence.

Step 3b is conditional on finding support for weak invariance (Dunbar et al., 2011). Metric equivalence would be indicated if a change of -.01 or less in the CFI fit index, a change of -.001 or less in the Gamma Hat fit index (Γ_1) and a change of -.02 or less in the McDonald Non-centrality index (Cheung & Rensvold, 2002) between the configural multi-group model and the weak invariance multi-group model is observed (Dunbar et al., 2011).

Step 4a: Establish whether the multi-group measurement model in which the structure of the model is constrained to be the same across groups, and in which all parameters are estimated freely across the samples, but for the factor loadings and the vector of regression intercepts, demonstrates acceptable fit when fitted to the samples simultaneously in a multi-group analysis.

The test of strong invariance determines whether the regression slopes and intercepts are the same across groups. A lack of strong invariance would imply that the intercepts of at least some of the items on the latent variable they represent differ across samples. Finding support for strong invariance would be a suitable result as it would support the position that the items operate in approximately the same way across samples in the way they reflect the underlying latent variables they are meant to reflect (Dunbar et al., 2011). If strong invariance (lack of uniform item bias) has been established, then scalar equivalence (step 4b) will be tested. If a lack of strong invariance (uniform item bias) has been established, then the process will be terminated[20] since, irrespective of what further tests might reveal, the measures are biased and observed score differences across genders cannot be interpreted to indicate corresponding differences in the latent variable.

Step 4b: Establish whether the multi-group measurement model in which the structure of the model is constrained to be the same across groups, and in which all parameters are estimated freely across the samples, but for the slope and the intercepts of the regression of the indicator variables on the latent variables, fits multi-group data poorer than a multi-group measurement model in which the structure of the model is constrained to be the same across groups but all parameters are estimated freely.

Step 4b is conditional on a finding of strong invariance (Dunbar et al., 2011). Scalar equivalence would be indicated if a change of -.01 or less in the CFI fit index, a change of -.001 or less in the Gamma Hat fit index (Γ_1) and a change of -.02 or less in the McDonald Non-centrality index (Cheung & Rensvold, 2002) between the configural multi-group model and the strong invariance multi-group model is observed (Dunbar et al., 2011).

[20] The process was again not terminated as further analyses were conducted in order to investigate partial strong invariance in order to determine which intercepts are biased. Biased intercepts were therefore identified and acknowledged in further tests of partial invariance and partial equivalence.

The test of scalar equivalence tests the hypothesis that the vector of item intercepts is invariant across groups. In the case where intercept differences are not due to biases but due to threshold differences that are based on known/expected group differences, which are not seen as undesirable, a test of scalar equivalence is not suitable (Vandenberg & Lance, 2000).

Step 5a: Establish whether the multi-group measurement model in which the structure of the model is constrained to be the same across groups, and in which all parameters are estimated freely across the samples, but for the factor loadings, the vector of regression intercepts and the measurement error variances of the indicator variables, demonstrates acceptable fit when fitted to the samples simultaneously in a multi-group analysis.

The test of strict invariance determines whether the regression slope, intercept and error variances of indicator variables are the same across groups. A lack of strict invariance would imply that the error variance of indicator variables of at least some of the items on the latent variable they represent, differ across samples. Strict invariance indicates that the respondents from the different gender groups respond to the instrument in such a manner that no significant variance exists across samples in terms of error terms associated with the indicator variable (Dunbar et al., 2011). If strict invariance (lack of item bias) has been established, then conditional probability equivalence will be tested. If lack of strict invariance has been established, then the process will be terminated[21] since, irrespective of what further tests might reveal, the measures are biased and observed score differences across genders cannot be interpreted to indicate corresponding differences in the latent variable.

Step 5b: Establish whether the multi-group measurement model in which the structure of the model is constrained to be the same across groups, and in which all parameters are estimated freely across the samples, but for the factor loadings, regression intercepts and measurement error variances of the indicator variables, fits multi-group data poorer than a multi-group measurement model in which the structure of the model is constrained to be the same across groups but all parameters are estimated freely.

Step 5b is conditional on a finding of strict invariance (Dunbar et al., 2011). Conditional probability would be indicated if a change of -.01 or less in the CFI fit index, a change of -.001 or less in the Gamma Hat fit index (Γ_1) and a change of - .02 or less in the McDonald Non-centrality index (Cheung

[21] This study further investigated partial strict invariance in order to determine which error variances of the indicator variables were biased, therefore the process was not terminated. Biased items were therefore identified and acknowledged in further tests of partial invariance and partial equivalence.

& Rensvold, 2002) between the configural multi-group model and the strict invariance multi-group model is observed (Dunbar et al., 2011).

Step 6a: Establish whether the multi-group measurement model in which the structure of the model is constrained to be the same across groups, and in which all parameters are constrained to be the same across the samples, demonstrates acceptable fit when fitted to the samples simultaneously in a multi-group analysis.

Given a finding of conditional probability equivalence the question is asked whether the latent variable variances and covariances are invariant across groups. According to Vandenberg and Lance (2000) the test of complete invariance determines whether the samples use "equivalent ranges of the construct continuum to respond to the indicators reflecting the construct" (p. 39). If the null hypothesis of close fit cannot be rejected, measurement invariance across samples is indicated.

Step 6b: Establish whether the multi-group measurement model in which the structure of the model is constrained to be the same across groups, and in which all parameters are constrained to be equal across the samples fits the multi-group data poorer than a multi-group measurement model in which the structure of the model is constrained to be the same across groups but all parameters are estimated freely.

Step 6b is conditional on a finding of complete invariance (Dunbar et al., 2011). Full equivalence would be indicated if a change of -.01 or less in the CFI fit index, a change of -.001 or less in the Gamma Hat fit index (Γ_1) and a change of -.02 or less in the McDonald Non-centrality index (Cheung & Rensvold, 2002) between the configural multi-group model and the complete invariance multi-group model is observed (Dunbar et al., 2011).

If complete measurement invariance and full measurement equivalence are supported, the model may be said to be equivalent and further tests would not be required[22] (Vandenberg & Lance, 2000). However, complete invariance and full measurement equivalence would not be tested if the preceding tests point to the source and extent of a lack of invariance and/or non-equivalence.

Vandenberg and Lance (2000) stated that the evaluation of model fit should be based on the chi-square test in conjunction with other practical fit indices. If the chi-square value is statistically non-significant, it supports a well-fitting model. However, the chi-square may be statistically significant even if there exist only minor differences between groups due to its sensitivity, especially to sample size. Vandenberg and Lance (2000) recommended that the following fit indices be used in

[22] This study continued investigations into partial complete invariance and partial full equivalence in order to identify biased items.

conjunction with the chi-square: (a) the Tucker-Lewis index (TLI) often referred to as the non-normed fit index (NNFI); (b) the relative non-centrality index (RNI) (c) the root mean square error of approximation (RMSEA), and (d) the standardised root mean square residual (SRMR).

The proposed testing procedure allows for the examination of measurement invariance and equivalence in a systematic way and if measurement instruments are invariant, cross-gender comparisons can be made. This has been found to be a sufficient testing procedure in order to ensure the comparability of the scores derived from psychological assessments across different gender groups. This section illustrated that measurement invariance and equivalence is much broader than measurement bias but includes all conceptualisations of measurement bias.

4.6 STATISTICAL POWER

Statistical power is important when making crucial decisions regarding the rejection or not of statistical hypotheses about model fit. Statistical power represents the probability that effects that actually exist have a chance of producing statistical significance in the data analysis (Tabachnick & Fidell, 2007).

In the case of this study, statistical hypotheses of close fit were formulated for the single group measurement model and for the configural invariance, weak invariance, strong invariance, strict invariance and complete invariance multi-group measurement models in terms of the parameter values attained for the Root Mean Squared Error of Approximation (RMSEA) fit index. In the context of this study statistical power refers to the conditional probability of rejecting the null hypothesis given that it is false (P [reject H_0: RMSEA $\leq$ .05)|H_0 false]). In the context of SEM, statistical power therefore refers to the probability of rejecting an incorrect model. If the null hypothesis of close fit (H_0: RMSEA $\leq$ .05) would not be rejected, the question that arises is whether this result is due to a lack of statistical power or whether it accurately reflects the true state of affairs. This concern increases as sample size decreases. If the decision not to reject the null hypothesis of close fit results under conditions of low power, it causes ambiguity because it is not clear whether the decision was due to the accuracy of the model or to the insensitivity of the test to detect specification errors in the model. The decision not to reject the null hypothesis of close fit would constitute convincing evidence on the merit of the model to the extent that it would be found that the statistical power of the evaluation of close fit had reasonably high power. Conversely, however, if the null hypothesis of close fit would be rejected under conditions of extremely high power it would create the fear that a reasonably accurate model had been rejected because of the extreme sensitivity of the test for minor specification errors in the model.

To calculate statistical power, the methodology of MacCallum et al., (1996) was implemented, using the model specification as indicated previously. Tabachnick and Fidell (2007) indicate that it is often desired that power levels of at least .80 are found prior to continuing with analyses (.80 would signal an 80 percent probability of achieving a significant result if an effect exists). In the case of evaluating the fit of a measurement model this would mean it would be desirable that the conditional probability of rejecting the close fit null hypothesis, given that the fit of the model is actually mediocre, should be at least .80. The MacCallum et al. (1996) SAS syntax has been translated into SPSS for the power calculation (Spangenberg & Theron, 2005). To derive power estimates for the test of close fit, the effect sizes of .05 and .08 were captured into the syntax (as the values of RMSEA under H_0 and H_a respectively), along with a statistical significance level of .05. In addition, the sample size and degrees of freedom, based on each respective model, were entered into the syntax. The power calculation specifications and results for each model and sample are reported in Table 4.2.

The power values displayed in Table 4.2 was taken into consideration when interpreting the results of the tests of close fit reported in chapter 5.

Table 4.2

Power calculations

Measurement model	Alpha	$RMSEA_0$	$RMSEA_a$	*N*	*df*	Power
Single-group: male sample	.05	.05	.08	460	1616	1.00000
Single-group: female sample	.05	.05	.08	460	1616	1.00000
Configural invariance	.05	.05	.08	460	3232	1.00000
Weak invariance	.05	.05	.08	460	3282	1.00000
Strong invariance	.05	.05	.08	460	3341	1.00000
Strict invariance	.05	.05	.08	460	3400	1.00000
Complete invariance	.05	.05	.08	460	3445	1.00000

CHAPTER 5
RESULTS

5.1 INTRODUCTION

As described in Chapters 3 and 4, this research aimed to determine whether the SWSI is able to measure the latent stress variables given its constitutive definition of the stress construct as it intends to and, if so, whether the latent stress variable is measured equivalently across gender. The final operational hypotheses were described in Chapter 4. This chapter aims to provide evidence that is used to decide on the validity of the operational hypotheses presented at the beginning of the previous chapter. However, prior to conducting the CFAs necessary to evaluate the measurement invariance and equivalence of the given instrument, some analyses were performed that provided insight into the instrument's functioning. These analyses (i.e., item and dimensionality analyses as well as discriminant validity for each gender sample) assisted in gaining understanding into the psychometric integrity of the indicator variables that represent the various latent variables.

The results are presented in the following order: (a) missing values, (b) item analyses, (c) dimensionality analyses, (d) multivariate normality, (e) CFA of the single-group occupational stress measurement model for each gender sample, (f) discriminant validity, (g) CFA of the multi-group occupational stress measurement model, and (h) corresponding measurement equivalence test.

5.2 MISSING VALUES

The PRELIS analyses revealed a missing values problem that needed to be attended to prior to conducting further analyses on the datasets. The assumptions listed by Mels (2007), as discussed in section 4.5.1.3 (e.g. type of variables, skewness), were considered in the decision regarding which method should be used to treat missing values. In this study the observed variables were measured on a scale comprising of 5 response options and only a limited number of missing values occurred on the items comprising the various subscales of the SWSI (i.e. less than 30% of the data constitutes missing values). Table 5.1 depicts the distribution of missing values across items. The observed variables could be described as excessively skewed. The PRELIS output for the male and female sample indicated, under skewness, that the z-scores were well above 1.96 and the p-values below .05, with the exception of item gen1 in the female sample. The histograms also showed negatively skewed distributions. However, two out of the three assumptions were met to allow for the use of multiple imputation as an appropriate method to treat the missing values. Missing values were

imputed using the PRELIS programme (Jöreskog & Sörbom, 1996b). The imputed sample retained all 460 cases in each gender sample.

Table 5.1

Number of missing values across items

	GEN1	GEN2	GEN3	GEN4	GEN5	GEN6	GEN7	GEN8
Male	0	0	0	0	0	0	0	0
Female	0	0	1	0	0	0	0	0
	GEN9	RA1	RA2	RA3	RA4	RA5	RA6	RA7
Male	0	0	0	0	0	0	0	0
Female	0	0	0	0	0	0	0	0
	REL8	REL9	REL10	REL11	REL12	REL13	REL14	REL15
Male	0	0	0	0	1	0	0	0
Female	0	0	0	0	0	0	0	0
	TE16	TE17	TE18	TE19	TE20	CA21	CA22	CA23
Male	1	0	0	0	0	0	0	0
Female	0	0	0	1	1	0	1	1
	CA24	CA25	JS26	JS27	JS28	JS29	LA30	LA31
Male	0	0	0	0	0	0	0	0
Female	0	0	1	0	1	0	1	0
	LA32	LA33	LA34	LA35	LA36	WH37	WH38	WH39
Male	0	0	0	0	0	0	0	0
Female	1	0	1	0	0	0	0	1
	WH40	WH41	WH42	WH43	WH44	WL45	WL46	WL47
Male	0	0	0	0	0	0	0	0
Female	0	0	0	0	1	0	0	0
	WL48	WL49	WL50					
Male	0	0	0					
Female	0	0	1					

GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home interface; WL = Workload.

5.3 ITEM ANALYSES

As described in the previous chapter, item analysis was conducted on each of the occupational stress sub-scales. Item analyses were conducted to investigate: (a) the reliability of indicators of each latent variable and (b) homogeneity of each sub-scale. Item analyses were conducted on each gender sample separately. The SPSS 19 Scale Reliability Procedure was used to analyse the sub-scale items.

Initially, problematic items identified through item statistics were flagged and discussed. Thereafter the homogeneity for each occupational stress sub-scale was evaluated.

5.3.1 SUB-SCALE RELIABILITIES

Sub-scale reliabilities for the male and female samples are reported in Table 5.2 and 5.3 respectively. All the sub-scales obtained Cronbach alpha values higher than a seemingly stringent cut-off value of .80[23].

Reliability coefficients were calculated for all the SWSI sub-scales before and after the imputation. The values were compared to determine if the imputation had an effect on the alpha values reported. Upon review of Table 5.2 (male sample) one small alpha value change was observed for Relationships (increase of .001). This would suggest that imputation affected the scale reliabilities in a trivial way.

Table 5.2

Reliability of the SWSI sub-scales for the male sample

		Pre-imputation				Post-imputation			
Sub-scale	**Number of items**	**Valid cases**	**Alpha**	**Mean**	**Variance**	**Valid cases**	**Alpha**	**Mean**	**Variance**
General Work Stress	9	460	.896	18.80	42.942	460	.896	18.7978	42.942
Role Ambiguity	7	460	.826	14.78	28.228	460	.826	14.7804	28.228
Relationships	8	459	.923	15.84	54.526	460	.924	15.8739	54.842
Tools and Equipment	5	459	.897	11.20	26.906	460	.897	11.2174	26.92
Career Advancement	5	460	.829	12.82	26.716	460	.829	12.8239	26.716
Job Security	4	460	.875	9.52	17.435	460	.875	9.5196	17.435
Lack of Autonomy	7	460	.856	16.46	37.935	460	.856	16.4565	37.935
Work/Home Interface	8	460	.839	15.55	35.603	460	.839	15.5478	35.603
Workload	6	460	.888	14.13	33.104	460	.888	14.1261	33.104

For the female sample (Table 5.3), a similar trend was observed in which trivial changes in alpha values for the General Work Stress (increase of .001), Work/Home interface (increase of .001) and Workload (decrease of .001) sub-scales emerged. This would suggest that imputation affected the sub-scale reliabilities in a trivial way.

[23] Even though Nunnally (1978, p. 245) indicates that "in the early stages of research on predictor tests or hypothesized measures of a construct, one saves time and energy by working with instruments that have only modest reliability, for which purpose reliabilities of 0.70 or higher will suffice" he nonetheless then continues and argues that "in many applied settings of 0.80 is not nearly high enough. Although not frequently quoted Nunnally (1978, p. 246) continues and claims that "in those applied settings where important decisions are made with respect to specific test scores, a reliability of 0.90 is the minimum that should be tolerated, and a reliability of 0.95 should be considered the desirable standard."

Table 5.3

Reliability of the SWSI sub-scales for the female sample

		Pre-imputation				Post-imputation			
Sub-scale	**Number of items**	**Valid cases**	**Alpha**	**Mean**	**Variance**	**Valid cases**	**Alpha**	**Mean**	**Variance**
General Work Stress	9	459	.908	20.19	49.674	460	.909	20.2087	49.734
Role Ambiguity	7	460	.833	15.08	31.028	460	.833	15.0848	31.028
Relationships	8	460	.933	15.96	67.177	460	.933	15.9587	67.177
Tools and Equipment	5	458	.913	10.97	29.839	460	.913	10.9609	29.763
Career Advancement	5	458	.873	12.16	31.126	460	.873	12.1652	31.079
Job Security	4	459	.890	9.85	19.911	460	.890	9.8652	19.925
Lack of Autonomy	7	457	.845	16.35	41.742	460	.845	16.4043	41.936
Work/Home Interface	8	458	.865	17.28	47.362	460	.866	17.2804	47.549
Workload	6	459	.879	13.89	34.175	460	.878	13.8826	34.108

Overall, the results of the reliability analyses would suggest satisfactory levels of internal consistency. In addition, the results suggested that the imputation of missing values has not affected the reliability results.

5.3.2 ITEM STATISTICS

Inspection of the means and standard deviations (i.e. for extreme means or small standard deviations) revealed no items, in each sub-scale over both samples that had to be flagged as problematic. Screening was based on the following item statistics: (a) inter-item correlations (b) corrected item-total correlations and squared multiple correlations, (c) the change in scale variance when deleting an item and (d) the change in the sub-scale Cronbach alpha when deleting an item. Inter-item correlations were considered low when r_{ij} dropped below .30. The corrected item-total correlations and the squared multiple correlations were considered problematic when values were substantially lower than those obtained by the bulk of the items in a specific sub-scale. The change in sub-scale variance when deleting an item was considered problematic when the sub-scale variance either increased upon deletion of an item or changed very little. The change in the sub-scale Cronbach alpha when deleting an item was considered problematic when the Cronbach alpha increased substantially well. Item statistics information is available in Appendix 1 on the accompanying CD.

For the male sample, within the Role Ambiguity sub-scale, items ra1 and ra3, ra2 and ra3, and ra3 and ra5 obtained inter-item correlations below .30. However, upon review of the item-total statistics, no items raised any cause for concern as they met all the cut-off values. The Tools and Equipment sub-scale indicated that if item te20 were deleted, the Cronbach Alpha value would

increase from .897 to .899. The other item statistics obtained for Item te20 did not indicate any other problems, therefore raising the question whether the small increase in Cronbach alpha when deleting this item would be significant and warranted. The item statistics for the sub-scale Lack of Autonomy showed that the squared multiple correlation for item la30 was below .30, but again, all other cut-off values for this item were met. The Work/Home interface results showed inter-item correlations below .30 (with regards to items wh37 and wh38, wh37 and wh43, and wh38 and wh40), but all other values for these items were above the critical cut-off values.

In the female sample, items ra1, ra2, and ra3 of the Role Ambiguity sub-scale obtained inter-item correlations below .30. However, these items as well as the other items in this sub-scale met all the other critical cut-off values. As with the male sample, the item statistic results for item te20 in the Tools and Equipment sub-scale indicated that if this item were to be deleted, the Cronbach Alpha would increase from .913 to .918. However, item te20 met all the other requirements, and therefore no further evidence to substantiate the deletion of this item was obtained. The item statistics for the Lack of Autonomy sub-scale showed that the squared multiple correlation for item la30 was below .30, but all other requirements for this item were met. Similarly, the Work/Home interface sub-scale results showed inter-item correlations between items wh37 and wh43 below .30, but all other values for these items were above the critical cut-off values. Consequently none of these items were flagged as particularly problematic items.

For both the male and female sample, no items seemed to cause significant problems even though some raised a few questions. However, due to the confirmatory nature of this study, the above mentioned items were retained for the subsequent CFAs.

5.4 DIMENSIONALITY ANALYSES

As discussed in Chapter 4, the uni-dimensionality assumption was tested for each of the occupational stress sub-scales, as each of the latent variable sub-scales was intended to reflect essentially one-dimensional sets of items. The analyses assisted in gaining an understanding of the item functioning per scale in the questionnaire. In order to determine uni-dimensionality, both the number of factors extracted, the associated factor loadings and the percentage of large residual correlations were considered. Should sub-scales fail the uni-dimensionality assumption, the possibility of meaningful factor fission was investigated. Therefore, the question was asked whether the extracted factors constitute meaningful subthemes within the original latent occupational stress dimension. In the case of sub-scales where the uni-dimensionality assumption was challenged, irrespective of whether meaningful factor fission occurred, the ability of a single factor to account for the observed inter-item correlation matrix was also investigated. This approach was taken to

investigate the magnitude of the factor loadings when a single factor (as per the a *priori* model) was forced and to examine the magnitude of the residual correlations.

The dimensionality analyses were conducted by subjecting each occupational stress sub-scale to an unrestricted principle axis factor analysis with oblique rotation. This analysis was performed on each of the 9 occupational stress sub-scales. The factor analysis was conducted separately for each gender sample.

SPSS 19 was used for the abovementioned analyses. All items were retained, following item analyses, due to the confirmatory nature of this study. The eigenvalue-greater-than-unity rule of thumb and the scree test was used to determine the number of factors to extract (Tabachnick & Fidell, 2007). Factor loadings were interpreted as follows: (a) .30 to .40 was considered to meet the minimal level for interpretation of structure, (b) .50 or greater was considered practically significant, and (c) loadings exceeding .70 were considered indicative of a well-defined structure (Hair et al., 2006). The practical significance value of .50 or greater was used as a benchmark for these analyses. An item with a loading of .50 would denote that 25 percent of the variance in the item is accounted for by the factor. The credibility of the extracted solution was evaluated by calculating the percentage of large residual correlations. Residual correlations larger than .05 were considered to be large. All the results of the dimensionality analyses are recorded in Appendix 2 on the accompanying CD.

5.4.1 DIMENSIONALITY ANALYSES RESULTS: MALE SAMPLE

The results of the principle axis factor analyses for the male sample are summarised in Table 5.4. Two of the nine sub-scales failed the uni-dimensionality test, however a further sub-scale fitted the one factor structure but with excessively large percentage of residuals, thereby indicating that a two factor solution actually provided a more credible solution. The affected sub-scales were: (a) Role Ambiguity, (b) Workload and (c) Work/Home Interface. Each sub-scale will be discussed in detail.

Table 5.4

Factor analysis results for the SWSI sub-scales: male sample

Scale	KMO	Bartlett's Test	% Variance explained	Min factor loading	Max factor loading	% Non-redundant residuals
General Work Stress (GEN)	.897	2050.797	Factor 1: 55.268	.597	.790	30.00%
Role Ambiguity (RA)	.795	1272.952	Factor 1: 49.883	.322	.851	33.00%
			Factor 2: 1.17.339	.256	.770	
			Single forced factor	.500	.753	66.00%
Relationships (REL)	.921	2463.371	Factor 1: 65.291	.623	.834	28.00%
Tools and Equipment (TE)	.881	1357.516	Factor 1: 71.115	.668	.864	0.00%
Career Advancement (CA)	.845	799.325	Factor 1: 59.797	.618	.835	0.00%
Job Security (JS)	.815	932.96	Factor 1: 72.874	.769	.830	0.00%
Lack of Autonomy (LA)	.864	1249.275	Factor 1: 53.938	.510	.737	28.00%
Work/Home Interface (WH)	.814	1422.392	Factor 1: 48.085	-.146	.797	28.00%
			Factor 2: 13.682	-.1023	.028	
			Single forced factor	.548	.744	64.00%
Workload (WL)	.849	1556.522	Factor 1: 64.318	.706	.817	80.00%

5.4.1.1. GENERAL WORK STRESS

For the correlation matrix to be factor analysable, the correlations in the correlation matrix should be larger than .30 and significant ($p < .05$). All the correlations were greater than .30 and all were significant ($p < .05$), indicating that the matrix was factor analysable. The Kaiser-Meyer-Olkin (KMO) is a measure of sampling adequacy and reflects the ratio of the sum of the squared inter-item correlations to the sum of the squared inter-item correlations plus the sum of the squared partial inter-item correlations, summed across all correlations. The correlation matrix is deemed factor analysable when the KMO approaches unity, or at least achieves a value bigger than .60 (Tabachnick & Fidell, 2007). A KMO value of .897 was obtained. This provides sufficient evidence that the General Work Stress sub-scale was factor analysable. The Bartlett's Test of Sphericity tests the null hypothesis that the correlation matrix is an identity matrix in the population (i.e., the diagonal contains 1's and all off-diagonal are zero's) (Tabachnick & Fidell, 2007). The Bartlett's Test of

Sphericity indicated that H_0 could be rejected ($p < .05$) providing further support that the matrix was factor analysable.

One factor was extracted, since only one factor obtained an eigenvalue greater than one. The scree plot also suggested that a single factor should be extracted. The factor matrix indicated that all the items satisfactorily loaded on the single extracted factor as all factor loadings were greater than .50. The resultant factor structure is shown in Table 5.5. Furthermore, 30% of the reproduced correlations were larger than .05, suggesting that the factor solution provided an adequate explanation for the observed inter-item correlation matrix. The uni-dimensionality assumption was thus corroborated.

Table 5.5

Rotated factor structure for the General Work Stress sub-scale

	Factor
	1
gen6	.790
gen2	.781
gen8	.769
gen5	.719
gen1	.701
gen4	.685
gen7	.654
gen3	.634
gen9	.597

5.4.1.2. ROLE AMBIGUITY

Item analyses results indicated a few low correlations among the items (< .30), however nothing significant that warranted the deletion of any of the items. All items were retained (due to the nature of this study) for this scale when the dimensionality analysis was run. The correlation matrix showed that not all correlations were larger than .30. Items ra1 and ra3 correlated .194, items ra1 and ra4 correlated .237, items ra2 and ra3 correlated .279 and items ra3 and ra5 correlated .281. All correlations were, however, significant ($p < .05$). The scale obtained a KMO of .795 and the Bartlett's Test of Sphericity allowed for the null hypothesis to be rejected, thus indicating that the correlation matrix was factor analysable.

The Role Ambiguity sub-scale was intended to reflect a single underlying factor. However, contrary to this intention, two factors had to be extracted to adequately explain the observed correlation

matrix. This was indicated by two factors obtaining eigenvalues greater than 1. The pattern matrix[24] is depicted in Table 5.6.

Table 5.6

Rotated factor structure for the Role Ambiguity sub-scale

	Factor	
	1	**2**
ra2	**.872**	-.045
ra5	**.804**	-.126
ra1	**.694**	.117
ra7	**.475**	.274
ra3	-.057	**.797**
ra4	.031	**.745**
ra6	.359	**.376**

The four items that loaded on the first factor all appear to refer to specificity of instructions given, whereas item ra3 and item ra4 clearly loaded on the second factor and seem to reflect expectations regarding the individual's duties (De Bruin & Taylor, 2006a). Item ra6 appears to have loaded on both factors (but does tend to lean to load on factor 2 for this sample). The factor fission obtained on this sub-scale makes substantive theoretical sense. Therefore, the two factors may be justified.

However, according to the design intentions of the SWSI, Role Ambiguity was treated as a single, undifferentiated latent variable. In order to determine how well the items of the Role Ambiguity sub-scale reflect a single underlying latent variable, the analysis was re-run by forcing the extraction of a single factor. The resultant single-factor factor structure is shown in Table 5.7. All items loaded onto the one factor with factors loadings equal to or larger than .50 which can be considered as satisfactory.

[24] The pattern matrix contains the partial regression coefficients when regressing the items on the correlated factors. The partial regression coefficients reflect the unique relationship between the item and each factor when holding the other factors in the rotated factor structure constant (Tabachnick & Fidell, 2007).

Table 5.7

Factor matrix when forcing the extraction of a single factor (Role Ambiguity)

	Factor
	1
ra2	.753
ra5	.749
ra7	.674
ra6	.637
ra1	.629
ra4	.550
ra3	.500

The residual correlations were computed for both the two-factor and the one-factor solution. For the two-factor solution, 33% of non-redundant residuals had absolute values greater than .05, which suggested that the rotated factor solution provided an adequate explanation for the observed inter-item correlation matrix. The one-factor solution, however, failed to provide a credible explanation in that 66% of the residual correlations were greater than .05.

5.4.1.3. RELATIONSHIPS

The correlation matrix indicated that all correlations were larger than .30 and all were significant ($p < .05$). The sub-scale obtained a KMO of .921 and the Bartlett's Test of Sphericity allowed for the null hypothesis to be rejected, thus there was strong evidence that the correlation matrix was factor analysable.

One factor obtained an eigenvalue greater than 1, therefore only one factor was extracted. This was further supported by the scree plot. The factor matrix indicated that all the items loaded satisfactorily on the single extracted factor as all factor loadings were larger than .50. The resultant factor structure is shown in Table 5.8. Furthermore, 28.0% of the reproduced correlations were larger than .05 suggesting that the factor solution provided a credible explanation for the observed inter-item correlation matrix. The uni-dimensionality assumption was thus supported.

Table 5.8

Rotated factor structure for the Relationships sub-scale

	Factor
	1
rel9	.834
rel10	.825
rel8	.813
rel11	.801
rel12	.794
rel14	.777
rel15	.735
rel13	.623

5.4.1.4. TOOLS AND EQUIPMENT

The results for this dimensionality analysis indicated that the correlation matrix was factor analysable as all the obtained correlations exceeded .30 and all were significant ($p < .05$). Furthermore, the KMO was .881 and the Bartlett's Test of Sphericity indicated that H_0 could be rejected.

Since only one factor obtained an eigenvalue greater than one, and upon inspection of the scree plot, one factor was extracted in terms of the observed correlation matrix. The factor matrix indicated satisfactory factor loadings ranging from .668 to .864. The resultant factor structure is shown in Table 5.9. Furthermore, 0% of the reproduced correlations were larger than .05. This suggests that the factor solution provides a credible explanation for the observed inter-item correlation matrix, and therefore the uni-dimensionality of the sub-scale was corroborated.

Table 5.9

Rotated factor structure for the Tools and Equipment sub-scale

	Factor
	1
te19	.864
te18	.841
te16	.830
te17	.790
te20	.668

5.4.1.5. CAREER ADVANCEMENT

Upon inspection of the dimensionality analysis performed on the Career Advancement sub-scale, all the items in the correlation matrix obtained correlations exceeding the .30 cut-off value and all were significant ($p < .05$). A KMO value of .845 was obtained for this sub-scale and the results indicated that the identity matrix H_0 could be rejected, meaning that the correlation matrix was factor analysable.

The results revealed that only one factor had to be extracted since only one factor obtained an eigenvalue greater than 1. The scree plot provided further evidence that only one factor should be extracted. The resultant factor structure is shown in Table 5.10. In terms of the proportion of variance that could be explained in each item by the first factor. All items could be regarded as satisfactory as all factor loadings were all larger than .50. Furthermore, 0% of the reproduced correlations were larger than .05. This suggested that the rotated factor solution provided a credible explanation for the observed inter-item correlation matrix, and therefore the uni-dimensionality of the sub-scale was corroborated.

Table 5.10

Rotated factor structure for the Career Advancement sub-scale

	Factor
	1
ca23	.835
ca21	.725
ca25	.701
ca24	.644
ca22	.618

5.4.1.6. JOB SECURITY

The correlation matrix showed that all correlations were larger than .30 and all were significant ($p < .05$). The sub-scale obtained a KMO of .815 and the Bartlett's Test of Sphericity allowed for the null hypothesis to be rejected ($p < .05$), thus there was strong evidence that the correlation matrix was factor analysable.

Only one factor obtained an eigenvalue greater than 1 and therefore one factor was extracted for the Job Security sub-scale. The scree plot also indicated the need to extract a single factor. The factor matrix, shown in Table 5.11, indicated that all the items loaded satisfactorily on the single

extracted factor (all factor loadings were larger than .50). Furthermore, 0% of the reproduced correlations were larger than .05 suggesting that the rotated factor solution provided a credible explanation for the observed inter-item correlation matrix. The uni-dimensionality assumption was thus corroborated.

Table 5.11

Rotated factor structure for the Job Security sub-scale

	Factor
	1
js28	.830
js29	.811
js27	.787
js26	.769

5.4.1.7. LACK OF AUTONOMY

The results for this dimensionality analysis indicated that the correlation matrix was factor analysable as all the correlations exceeded .30 and all were significant ($p < .05$). The KMO for this sub-scale was .864 and the Bartlett's Test of Sphericity indicated that the identity matrix H_0 could be rejected.

One factor was extracted in terms of the observed correlation matrix, since only one factor obtained an eigenvalue greater than 1. The factor matrix indicated that all the items satisfactorily loaded onto the single extracted factor. All the obtained factor loadings were bigger than .50 and only 28.0% of the reproduced correlations were larger than .05, suggesting that the solution provided a credible explanation for the observed inter-item correlation matrix. The resultant factor structure is shown in Table 5.12. The uni-dimensionality assumption for this sub-scale was thus verified.

Table 5.12

Rotated factor structure for the Lack of Autonomy sub-scale

	Factor
	1
la35	.737
la33	.735
la36	.703
la31	.703
la34	.682
la32	.676
la30	.510

5.4.1.8. WORK/HOME INTERFACE

For this sub-scale the item analyses results indicated a few low correlations among the items (< .30), however nothing significant to warrant the deletion of any of the items. Due to the confirmatory nature of this study, all items were retained for this sub-scale when the dimensionality analysis was run. The correlation matrix showed that a few correlations were smaller than .30. Items wh37 and wh38 correlated .286, items wh37 and wh43 correlated .240, items wh38 and wh40 correlated .286, items wh38 and wh44 correlated .190 and items wh42 and wh44 correlated .281. All correlations were, however, significant ($p < .05$). The sub-scale obtained a KMO of .814 and the Bartlett's Test of Sphericity allowed for the null hypothesis to be rejected ($p < .05$), thus indicating that the correlation matrix was factor analysable.

The items of the Work/Home Interface sub-scale were intended to reflect a single underlying factor. However, contrary to this intention, two factors had to be extracted to adequately explain the observed correlation matrix. This was indicated by two factors obtaining eigenvalues greater than 1. The scree plot was somewhat ambiguous as to whether one or two factors should be extracted. The pattern matrix is depicted in Table 5.13.

Table 5.13

Rotated factor structure for the Work/Home Interface sub-scale

	Factor	
	1	**2**
wh41	**.797**	.024
wh42	**.690**	.028
wh43	**.609**	.026
wh40	**.608**	-.135
wh38	**.548**	-.043
wh37	-.146	**-1.023**
wh39	.221	**-.643**
wh44	.183	**-.470**

The five items that loaded on the first factor all appear to refer to the home aspect of this sub-scale, whether it being the lack of support from home or the negative effects of happenings at home on work or happenings at work on home. The three items that loaded on the second factor seemed to reflect more of a balance between the individual's work and home life. The factor fission obtained on this sub-scale makes substantive theoretical sense. Therefore, the two factors may be justified.

However, according to the design intentions of the SWSI, Work/Home Interface was treated as a single, undifferentiated latent variable. Upon forcing the extraction of a single factor solution on the data, all the items loaded satisfactorily on the single factor with all factors loadings larger than .50. The resultant single-factor factor structure is shown in Table 5.14.

Table 5.14

Factor matrix when forcing the extraction of a single factor (Work/Home Interface)

	Factor
	1
wh39	.744
wh41	.714
wh40	.691
wh37	.647
wh42	.617
wh44	.571
wh38	.553
wh43	.548

The residual correlations were computed for both the two-factor and the one-factor solution. For the two-factor solution, 28% of non-redundant residuals had absolute values greater than .05. This can still suggest that the rotated factor solution is an adequate explanation for the observed inter-

item correlation matrix. The one-factor solution, however, failed to provide a credible explanation in that 64% of the residual correlations were greater than .05.

5.4.1.9. WORKLOAD

The correlation matrix indicated that all correlations were larger than .30 and all were significant ($p < .05$). The sub-scale obtained a KMO of .849 and the Bartlett's Test of Sphericity allowed for the identity matrix null hypothesis to be rejected. This provided strong evidence that the correlation matrix was factor analysable.

One factor obtained an eigenvalue greater than 1, therefore only one factor was extracted. This was further supported by the scree plot. The factor matrix indicated that all the items satisfactorily loaded on the one extracted factor as all factor loadings were larger than .70. The resultant factor structure is shown in Table 5.15. However, 80.0% of the reproduced correlations were larger than.05 suggesting that the rotated factor solution fails to provide a credible explanation for the observed inter-item correlation matrix.

Table 5.15

Rotated factor structure for the Workload sub-scale

	Factor
	1
wl49	.817
wl48	.797
wl47	.756
wl46	.736
wl50	.722
wl45	.706

The large percentage of large residuals suggests the possibility of a second factor. The eigenvalue associated with the second factor was .804, which only marginally missed the Kaiser criterion for extraction. When a second factor was forced nd the solution rotated to simple structure, Table 5.16 indicates that a meaningful structure does emerge. Three items loaded satisfactorily on factor 1 and three items satisfactorily loaded negatively on factor 2.

Table 5.16

Pattern matrix when forcing the extraction of two factors (Workload)

	Factor	
	1	**2**
wl49	**.957**	.051
wl50	**.769**	-.011
wl45	**.674**	-.080
wl47	-.082	**-.933**
wl48	.032	**-.846**
wl46	.215	**-.568**

The residual correlations, computed for the two-factor solution, indicated that 0% of non-redundant residuals had absolute values greater than .05. This suggests that the two-factor solution is a more credible explanation for the observed inter-item correlation matrix.

5.4.2 DIMENSIONALITY ANALYSES RESULTS: FEMALE SAMPLE

The results of the principle axis factor analyses for the female sample are summarised in Table 5.17. Only one of the nine sub-scales failed the uni-dimensionality test, however three sub-scales fitted the one factor structure but with excessively large percentage of residuals, indicating that a two-factor solution actually provides a more credible solution. The affected sub-scales were: (a) Role Ambiguity, (b) Job Security, (c) Workload and (d) Work/Home Interface. The results for each sub-scale will be discussed in detail.

Table 5.17

Factor analysis results for the SWSI scales: female sample

Scale	KMO	Bartlett's Test	% Variance explained	Min factor loading	Max factor loading	% Non-redundant residuals
General Work Stress (GEN)	.915	2254.94	58.061	.666	.792	41.00%
Role Ambiguity (RA)	.792	1453.451	Factor 1: 50.797	-.001	.911	4.00%
			Factor 2: 19.260	-.077	.841	
			Single forced factor	.450	.779	71.00%
Relationships (REL)	.929	2681.188	Factor 1: 68.073	.692	.874	35.00%
Tools and Equipment (TE)	.857	1656.606	Factor 1: 74.609	.691	.886	30.00%
Career Advancement (CA)	.873	1081.355	Factor 1: 66.418	.691	.837	0.00%
Job Security (JS)	.792	1143.99	Factor 1: 75.442	.736	.910	50.00%
Lack of Autonomy (LA)	.873	1166.296	Factor 1: 52.964	.562	.768	42.00%
Work/Home Interface (WH)	.849	1606.642	Factor 1: 52.007	.515	.785	50.00%
Workload (WL)	.831	1503.641	Factor 1: 62.302	.676	.781	80.00%

5.4.2.1. GENERAL WORK STRESS

For the correlation matrix to be factor analysable, the correlations in the correlation matrix should be larger than .30 and significant ($p < .05$). All the correlations were greater than .30 and all were significant ($p < .50$), indicating that the matrix of the General Work Stress sub-scale for the female sample was factor analysable. A KMO value of .915 was obtained. This provides sufficient evidence that the General Work Stress sub-scale was factor analysable (KMO > .60). The Bartlett's Test of Sphericity indicated that H_0 could be reject ($p < .05$) providing further support that the matrix was factor analysable.

One factor was extracted, since only one factor obtained an eigenvalue greater than one. The scree plot also suggested that a single factor should be extracted. The factor matrix indicated that all the items satisfactorily loaded on the one extracted factor as all factor loadings were greater than .50. The resultant factor structure is shown in Table 5.18. Furthermore, 41% of the reproduced correlations were larger than .05, suggesting that the rotated factor solution provides a somewhat

tenuous explanation for the observed inter-item correlation matrix. The uni-dimensionality assumption was nonetheless considered corroborated.

Table 5.18

Rotated factor structure for the General Work Stress sub-scale

	Factor
	1
gen2	.792
gen6	.757
gen8	.745
gen3	.724
gen1	.723
gen4	.716
gen7	.714
gen5	.700
gen9	.666

5.4.2.1. ROLE AMBIGUITY

Item analyses results for the Role Ambiguity sub-scale indicated a few low correlations among the items (< .30), however nothing significant to warrant the deletion of any of the items. All items were retained (due to the nature of this study) for this sub-scale when the dimensionality analysis was run. The correlation matrix indicated a few correlations smaller than .30. Items ra1 and ra3 correlated .208, ra1 and ra4 correlated .193, items ra2 and ra4 correlated .210, items ra2 and ra3 correlated .253 and items ra3 and ra5 correlated .251. All correlations were, however, significant ($p < .05$). The scale obtained a KMO of .792 and the Bartlett's Test of Sphericity allowed for the null hypothesis to be rejected, thus indicating that the correlation matrix was factor analysable.

The design intention was that all the items of the Role Ambiguity sub-scale should reflect a single underlying factor. However, contrary to this intention, two factors had to be extracted to adequately explain the observed correlation matrix. This was indicated by two factors obtaining eigenvalues greater than 1. A similar finding was obtained for the male sample. The pattern matrix is depicted in Table 5.19.

Table 5.19

Rotated factor structure for the Role Ambiguity sub-scale

	Factor	
	1	**2**
ra1	**.911**	-.133
ra2	**.889**	-.077
ra5	**.756**	.068
ra7	**.546**	.254
ra6	**.358**	.343
ra4	-.037	**.841**
ra3	-.001	**.739**

The five items that loaded on the first factor all appeared to refer to specificity of instructions given. The two items loading on the second factor seemed to reflect expectations regarding the individual's duties. As with the male sample, item ra6 loaded similarly on both factors. However, with this female sample, item ra6 loaded slightly higher on factor 1 as opposed to the male sample where it loaded slightly higher on factor 2. The factor fission obtained on this sub-scale makes substantive theoretical sense. Therefore, the two factors may be justified.

However, according to the design intentions of the SWSI, Role Ambiguity was treated as a single, undifferentiated latent variable. In order to determine how well the items of the Role Ambiguity sub-scale reflect a single underlying latent variable, the analysis was re-run by forcing the extraction of a single factor. The resultant single-factor factor structure is shown in Table 5.20. Five of the seven items loaded onto the one factor with factor loadings equal to or larger than .50, which can be considered as satisfactory. However, items ra4 and ra3 obtained loadings that fell below the cut-off value set for this study.

Table 5.20

Factor matrix when forcing the extraction of a single factor (Role Ambiguity)

	Factor
	1
ra2	.779
ra5	.779
ra1	.751
ra7	.713
ra6	.582
ra4	.465
ra3	.450

The residual correlations were computed for both the two-factor and the one-factor solution. For the two-factor solution, only 4.0% of non-redundant residuals had absolute values greater than .05. This suggests that the rotated factor solution is an adequate explanation for the observed inter-item correlation matrix. The one-factor solution, however, as could be expected, failed to provide a credible explanation in that 71.0% of the residual correlations were greater than .05.

5.4.2.2. RELATIONSHIPS

The correlation matrix indicated that all correlations were larger than .30 and all were significant ($p < .05$). The sub-scale obtained a KMO of .929 and the Bartlett's Test of Sphericity allowed for the null hypothesis to be rejected, thus there was strong evidence that the correlation matrix was factor analysable.

One factor obtained an eigenvalue greater than 1, therefore only one factor was extracted. This was further supported by the scree plot. The factor matrix indicated that all the items satisfactorily loaded on the one extracted factor as all factor loadings were larger than .50. The resultant factor structure is shown in Table 5.21. Furthermore only 35.0% of the reproduced correlations were larger than .05 suggesting that the rotated factor solution provided a reasonably credible explanation for the observed inter-item correlation matrix. The uni-dimensionality assumption was thus supported.

Table 5.21

Rotated factor structure for the Relationships sub-scale

	Factor
	1
rel12	.874
rel11	.827
rel10	.819
rel14	.797
rel8	.792
rel15	.786
rel9	.782
rel13	.692

5.4.2.3. TOOLS AND EQUIPMENT

The results for this dimensionality analysis indicated that the correlation matrix was factor analysable as all the obtained correlations exceeded .30 and all were significant ($p < .05$). Furthermore, the KMO was .857 and the Bartlett's Test of Sphericity indicated that H_0 could be rejected.

Since only one factor obtained an eigenvalue greater than one, and upon inspection of the scree plot, one factor was extracted to account for the observed correlation matrix. The factor matrix indicated satisfactory factor loadings ranging from .691 to .886. The resultant factor structure is shown in Table 5.22. Furthermore, 30.0% of the reproduced correlations were larger than .05. This suggested that the rotated factor solution provided a credible explanation for the observed inter-item correlation matrix, and therefore the uni-dimensionality was corroborated.

Table 5.22

Rotated factor structure for the Tools and Equipment sub-scale

	Factor
	1
te19	.886
te18	.875
te17	.85
te16	.825
te20	.691

5.4.2.4. CAREER ADVANCEMENT

Upon inspection of the results of the dimensionality analysis performed on the Career Advancement sub-scale, all the items in the correlation matrix obtained correlations exceeding the .30 cut-off value and were all significant ($p < .05$). A KMO value of .873 was obtained for this sub-scale and the results indicated that the identity matrix H_0 could be rejected, meaning that the correlation matrix was factor analysable.

The results revealed that only one factor had to be extracted since only one factor obtained an eigenvalue greater than 1. The scree plot provided further evidence that only one factor needed to be extracted. The resultant factor structure is shown in Table 5.23. In terms of the proportion of variance in each item that could be explained by the extracted factor, all item factor loadings could be regarded as satisfactory as they were all larger than .50. Furthermore, 0% of the reproduced correlations were larger than .05. This suggested that the factor solution provided a credible explanation for the observed inter-item correlation matrix, and therefore the uni-dimensionality was corroborated.

Table 5.23

Rotated factor structure for the Career Advancement sub-scale

	Factor
	1
ca21	.837
ca23	.827
ca25	.759
ca22	.692
ca24	.691

5.4.2.5. JOB SECURITY

The correlation matrix showed that all correlations were larger than .30 and all were significant ($p < .05$). The sub-scale obtained a KMO of .792 and the Bartlett's Test of Sphericity allowed for the null hypothesis to be rejected ($p < .05$), thus there was strong evidence that the correlation matrix was factor analysable.

Only one factor obtained an eigenvalue greater than 1 and the elbow in the scree plot further supported the need to extract one factor for the Job Security sub-scale. The factor matrix, shown in Table 5.24, indicated that all the items satisfactorily loaded on the single extracted factor as all factor loadings were larger than .50, ranging from .736 to .910. However, 50.0% of the reproduced

correlations were larger than .05 suggesting that the factor solution may not provide a credible explanation for the observed inter-item correlation matrix.

Table 5.24

Rotated factor structure for the Job Security sub-scale

	Factor
	1
js27	.910
js28	.868
js29	.765
js26	.736

The large percentage of large residuals suggested the possibility of a second factor. The eigenvalue associated with the second factor was .527, which noticeably missed the Kaiser criterion for extraction. However, when a second factor was forced and the solution rotated to simple structure, Table 5.25 indicates that a meaningful structure does emerge. Two items loaded satisfactorily on factor 1 and factor 2, respectively.

Table 5.25

Pattern matrix when forcing the extraction of two factors (Job Security)

	Factor	
	1	**2**
js29	**.883**	-.051
js28	**.741**	.173
js26	-.050	**.862**
js27	.251	**.710**

The residual correlations, computed for the two-factor solution, indicated that 0% of non-redundant residuals had absolute values greater than .05. This suggested that the two-factor solution was a credible explanation for the observed inter-item correlation matrix.

5.4.2.6. LACK OF AUTONOMY

All the correlations in the correlation matrix exceeded .30 except for items la30 and la36, which correlated .283. However, all correlations were significant ($p < .05$). The KMO for this sub-scale was .873 and the Bartlett's Test of Sphericity indicated that H_0 could be rejected. This provided strong evidence that the correlation matrix was factor analysable.

One factor was extracted to explain the observed correlation matrix, since only one factor obtained an eigenvalue greater than 1. The factor matrix indicated that all the items loaded satisfactorily onto the one extracted factor. All the obtained factor loadings were larger than .50. Furthermore, 42.0% of the reproduced correlations were larger than .05, suggesting that the solution provided a somewhat unconvincing explanation for the observed inter-item correlation matrix. The resultant factor structure is shown in Table 5.26.

Table 5.26

Rotated factor structure for the Lack of Autonomy sub-scale

	Factor
	1
la33	.768
la34	.750
la32	.708
la31	.663
la36	.661
la35	.578
la30	.562

5.4.2.7. WORK/HOME INTERFACE

The results of the item analyses for this sub-scale indicated a few low correlations among the items (< .30), however nothing significant to warrant the deletion of any of the items. Due to the confirmatory nature of this study, all items were retained for this sub-scale when the dimensionality analysis was run. The correlation matrix showed that a few correlations were smaller than .30. Items wh37 and wh43 correlated .271, and items wh39 and wh43 correlated .261. All correlations were, however, significant ($p < .05$). The sub-scale obtained a KMO of .849 and the Bartlett's Test of Sphericity allowed for the null hypothesis to be rejected ($p < .05$), thus indicating that the correlation matrix was factor analysable.

The Work/Home Interface sub-scale was intended to consist of a one-dimensional set of items. Contrary to the results obtained when dimensionality analysis was run for this sub-scale with the male sample, one factor was extracted in terms of the observed correlation matrix for the female sample. Only one factor obtained an eigenvalue greater than 1. The factor matrix indicated that all the items loaded satisfactorily onto the one extracted factor. All the obtained factor loadings were larger than .50. However, 50.0% of the reproduced correlations were larger than .05, suggesting that

the one-factor solution failed to provide a convincing explanation for the observed inter-item correlation matrix. The resultant factor structure is shown in Table 5.27.

Table 5.27

Rotated factor structure for the Work/Home Interface sub-scale

	Factor
	1
wh39	.785
wh40	.737
wh41	.703
wh37	.692
wh42	.668
wh44	.653
wh38	.604
wh43	.515

The large percentage of large residuals suggests the possibility of a second factor. The eigenvalue associated with the second factor was .980, which only marginally missed the Kaiser criterion for extraction. When a second factor was forced and the solution rotated to simple structure, Table 5.28 indicates that a meaningful structure does emerge. Four items loaded on factor 1, with one item below .50, and four items loaded negatively on factor 2, with two items below .50. The Work/Home interface sub-scale has a clear splitting of the items. Half the items refers to the home aspect of the sub-scale, whether it being the lack of support from home or the negative effects of happenings at home on work or happenings at work on home. The other half of the items refer to a balance between the individual's work and home life.

Table 5.28

Pattern matrix when forcing the extraction of two factors (Work/Home interface)

	Factor	
	1	2
wh41	**.834**	.034
wh40	**.682**	-.131
wh43	**.640**	.067
wh42	**.488**	-.236
wh39	-.078	**-.993**
wh37	-.009	**-.785**
wh38	.233	**-.422**
wh44	.303	**-.404**

The residual correlations, computed for the two-factor solution, indicated that 28.0% of non-redundant residuals had absolute values greater than .05. This suggested that the two-factor

solution provided a more adequate explanation for the observed inter-item correlation matrix as opposed to the one-factor solution in Table 5.27.

5.4.2.8. WORKLOAD

The correlation matrix indicated that all correlations were larger than .30 and all were significant ($p < .05$). The sub-scale obtained a KMO of .831 and the Bartlett's Test of Sphericity allowed for the null hypothesis to be rejected. This provided strong evidence that the correlation matrix was factor analysable.

One factor obtained an eigenvalue greater than 1, therefore only one factor was extracted. This was further supported by the scree plot. The factor matrix indicated that all the items satisfactorily loaded on the single extracted factor as all factor loadings were larger than .50. The resultant factor structure is shown in Table 5.29. However, 80.0% of the reproduced correlations were larger than .05, strongly suggesting that the single factor solution failed to provide a credible explanation for the observed inter-item correlation matrix.

Table 5.29

Rotated factor structure for the Workload sub-scale

	Factor
	1
wl48	.781
wl49	.770
wl46	.765
wl47	.744
wl50	.701
wl45	.676

The large percentage of large residuals suggested the possibility of a second factor. The eigenvalue associated with the second factor was .935, which only marginally missed the Kaiser criterion for extraction. When a second factor was forced and the solution rotated to simple structure, Table 5.30 indicates that a meaningful structure did emerge. Three items loaded satisfactorily on factor 1 and three items obtained satisfactory loadings (negative) on factor 2.

Table 5.30

Pattern matrix when forcing the extraction of two factors (Workload)

	Factor	
	1	**2**
wl50	**.814**	.021
wl49	**.789**	-.070
wl45	**.753**	-.003
wl47	-.142	**-1.046**
wl48	.150	**-.716**
wl46	.254	**-.579**

The residual correlations, computed for the two-factor solution, indicated that 0% of non-redundant residuals had absolute values greater than .05. This suggested that the two-factor solution was a more credible explanation for the observed inter-item correlation matrix.

5.5 CONCLUSIONS DERIVED FROM THE ITEM AND DIMENSIONALITY ANALYSES

The purpose of the foregoing analyses was to provide insight into the functioning of the SWSI occupational stress sub-scales. Further to this, the analyses assisted in gaining an understanding about the psychometric integrity of the indicator variables that were tasked to represent each of the occupational stress latent variables.

The item analyses revealed that sufficient internal consistency for the SWSI occupational sub-scales was established. In all cases, the sub-scales achieved alpha values exceeding .80 in both samples. This provided evidence in support of the homogeneity of each sub-scale as proposed by the test publishers. At a more detailed level, the item correlations and statistics revealed few items that were flagged as being potentially problematic. However, overall the items performed well for each gender sample.

As far as the dimensionality analyses are concerned, only two of the nine sub-scales (Role Ambiguity and Work/Home Interface) for the male sample and only one of the nine sub-scales (Role Ambiguity) for the female sample did not formally satisfy the uni-dimensionality assumption. When the credibility of the solution extracted based on the Kaiser criterion was also taken into account, a further sub-scale (Workload) failed to satisfy the uni-dimensionality assumption in the male sample and a further three sub-scales (Job Security, Work/Home Interface and Workload) in the female sample. The items of the sub-scales (Role Ambiguity and Work/Home Interface) that failed to satisfy the uni-dimensionality assumption (i.e. Role Ambiguity sub-scale and Work/Home Interface sub-

scale in the male sample and Role Ambiguity sub-scale in the female sample) were successfully forced onto a single factor solution. However, the residuals calculated from the inter-item correlation matrix and the reproduced matrix indicated that the initial solution, prior to forcing a single factor, provided a more convincing explanation for the observed inter-item correlation matrix. This suggested that these factors could be better explained by further sub-facets of the respective occupational stress sub-dimensions.

Furthermore, four of the sub-scales (the Workload sub-scale in the male sample and the Job Security, Work/Home Interface, and Workload sub-scales in the female sample) reflected a large percentage of large residuals, thereby failing to provide a credible explanation for the observed inter-item correlation matrix. When a second factor was forced, a meaningful and credible factor structure emerged, suggesting that these sub-scales could be better explained by further sub-facets of the respective occupational stress sub-dimensions.

5.6 MULTIVARIATE NORMALITY

Multivariate statistics in general, and structural equation modelling in particular, are based on a number of critical assumptions. It was necessary to assess the extent to which the data complies with these assumptions before proceeding with the main analyses (Tabachnick & Fidell, 2007). The default method of estimation when fitting measurement models to continuous data, maximum likelihood, assumes that the distribution of indicator variables follow a multivariate normal distribution (Mels, 2003). If this assumption is not met, the results will include incorrect standard errors and chi-square estimates (Du Toit & Du Toit, 2001; Mels, 2003).

The multivariate normality of the multivariate item distribution in this study was evaluated via PRELIS. The results of the test of multivariate normality of the SWSI multivariate item distributions are depicted in Table 5.31. Detailed results of the tests of multivariate normality for continuous variables, before and after normalisation for both the male and females sample has been reported in Appendix 3 on the accompanying CD.

Table 5.31

Tests of multivariate normality for continuous variables: before normalisation

		Skewness			Kurtosis			Skewness and Kurtosis	
Sample	**N**	**Value**	**Z-Score**	**P-Value**	**Value**	**Z-Score**	**P-Value**	**Chi-Square**	**P-Value**
Male	460	758.525	69.806	0.000	4193.044	29.508	0.000	5743.552	0.000
Female	460	877.874	93.377	0.000	4371.501	32.806	0.000	9795.541	0.000

The chi-square values for skewness and kurtosis indicated that all individual indicator variables failed the test of univariate normality ($p < .05$) in the male sample (univariate results not included in the table). Only one of the fifty individual indicator variables (gen1) did not fail the test of univariate normality ($p > .05$) in the female sample. Furthermore, the null hypothesis that the data follows a multivariate normal distribution also had to be rejected for the male and female sample (χ^2 = 5743.552; $p < .05$ and χ^2 = 9795.541; $p < 0.05$ respectively). It was decided to attempt normalising both the datasets with PRELIS. The results of the test for multivariate normality on the normalised indicator variables (only multivariate results) are presented in Table 5.32.

Table 5.32

Tests of multivariate normality for continuous variables: after normalisation

		Skewness			Kurtosis			Skewness and Kurtosis	
Sample	**N**	**Value**	**Z-Score**	**P-Value**	**Value**	**Z-Score**	**P-Value**	**Chi-Square**	**P-Value**
Males	460	698.022	56.900	0.000	4093.211	27.215	0.000	3978.295	0.000
Females	460	747.751	67.559	0.000	4183.937	29.315	0.000	5423.556	0.000

Normalising the data typically does improve the symmetry and kurtosis of the individual indicator variable distributions. The chi-square values for skewness and kurtosis (not shown) indicated that the null hypothesis of univariate normality need not be rejected for the individual items ($p > .05$) in both samples. The decrease in the chi-squared value for each sample in Table 5.32 indicated that although the attempt at normalising the data was not successful ($p < 0.05$), it did improve the situation (a decrease from 5743.552 to 3978.295 for the male sample and a decrease from 9795.541 to 5423.556 for the female sample). Due to the above findings indicating that the normalisation

option did not have the desired effect, the use of an alternative method of estimation more suited to data not following a multivariate normal distribution was rather considered. As recommended by Mels (2003), Robust Maximum Likelihood (RML) was selected as the preferred estimation method for this study. The computation of an asymptotic covariance matrix via PRELIS was necessary to enable the calculation of more appropriate fit indices in LISREL. For this purpose the normalised data set was utilised due to the beneficial effect that the attempt at normalising the data had on the multivariate indicator variable distribution.

5.7 EVALUATING THE SWSI SINGLE-GROUP MEASUREMENT MODEL FIT VIA CONFIRMATORY FACTOR ANALYSIS IN LISREL

The relationship between the occupational stress latent variable and its manifest indicators is represented by the measurement model. The success with which the latent variables were operationalised in terms of the individual items is ultimately determined via confirmatory factor analysis. If the measurement model can successfully reproduce the observed covariance matrix (i.e., if the model fits well), and if the measurement model parameter estimates indicate that the majority of the variance in the indicator variables can be explained in terms of the latent variables they were tasked to reflect, then the operationalization can be considered successful.

As indicated by the operational hypotheses, the single-group occupational stress measurement model was fitted to each gender sample independently. This analysis specifically aimed to evaluate whether the single-group occupational stress measurement model implied by the scoring key of the SWSI could closely reproduce the covariances observed between the individual items comprising each of the occupational stress scales in the separate gender groups (operational hypothesis 1). The results of the fitted single-group measurement model for each gender group are recorded in Appendix 3 on the accompanying CD.

5.7.1 SINGLE-GROUP MEASUREMENT MODEL FIT: MALE SAMPLE

Operational hypothesis 1a was tested by testing H_{01a_male}: RMSEA ≤ .05. The fit of the occupational stress measurement model for the male sample was evaluated based on the array of model fit indices reported by LISREL. The standardised residuals, modification indices, as well as the measurement model parameter estimates were also evaluated. The discriminant validity of the inferences derived from the items earmarked to reflect specific stress dimensions was finally evaluated.

A visual representation of the SWSI measurement model that was fitted to the data of the male sample (*N*=460) is shown in Figure 5.1.

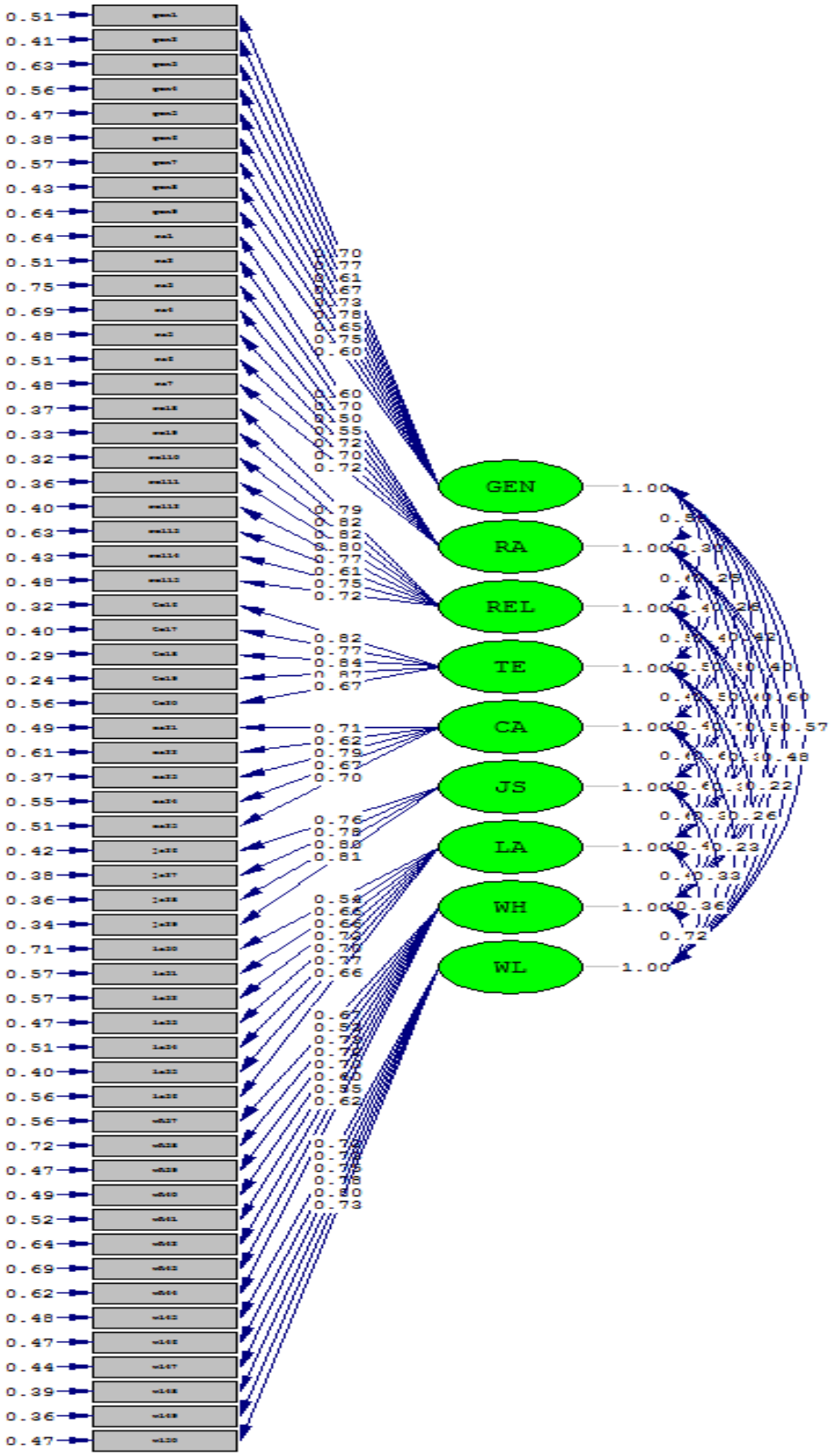

*Figure 5.1.*Representation of the fitted SWSI measurement model: male sample

5.7.1.1 MEASUREMENT MODEL FIT INDICES

The measurement model converged in 19 iterations. The full spectrum of fit statistics produced by LISREL 9.00 is shown in Table 5.33.

Table 5.33

Goodness of fit statistics for the SWSI measurement model: male sample

Degrees of Freedom = 1616
Minimum Fit Function Chi-Square = 3927.929 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 4227.187 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 3585.946 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 1969.946
90 Percent Confidence Interval for NCP = (1800.332 ; 2147.212)

Minimum Fit Function Value = 8.558
Population Discrepancy Function Value (F0) = 4.292
90 Percent Confidence Interval for F0 = (3.922 ; 4.678)
Root Mean Square Error of Approximation (RMSEA) = 0.0515
90 Percent Confidence Interval for RMSEA = (0.0493 ; 0.0538)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000

Expected Cross-Validation Index (ECVI) = 8.484
90 Percent Confidence Interval for ECVI = (8.114 ; 8.870)
ECVI for Saturated Model = 7.712
ECVI for Independence Model = 152.522

Chi-Square for Independence Model with 1711 Degrees of Freedom = 69889.784
Independence AIC = 70007.784
Model AIC = 3893.946
Saturated AIC = 3540.000
Independence CAIC = 70310.526
Model CAIC = 4684.155
Saturated CAIC = 12622.271

Normed Fit Index (NFI) = 0.949
Non-Normed Fit Index (NNFI) = 0.969
Parsimony Normed Fit Index (PNFI) = 0.896
Comparative Fit Index (CFI) = 0.971
Incremental Fit Index (IFI) = 0.971
Relative Fit Index (RFI) = 0.946

Critical N (CN) = 225.152

Root Mean Square Residual (RMR) = 0.0803
Standardized RMR = 0.0616
Goodness of Fit Index (GFI) = 0.762
Adjusted Goodness of Fit Index (AGFI) = 0.739
Parsimony Goodness of Fit Index (PGFI) = 0.696

Upon first inspection of Table 5.33, the degrees of freedom are indicated to be 1616. This corresponds with the earlier calculations (section 4.5.1.2) and provides evidence that the model was specified correctly.

The exact fit null hypothesis was not tested, as this represents a somewhat unrealistic position that the single-group measurement model is able to reproduce the observed covariance matrix to a degree of accuracy that could be explained in terms of sampling error only.

The following close fit null hypothesis was tested:

$$H_{01a\_male}: \text{RMSEA} \leq .05$$

$$H_{a1a\_male}: \text{RMSEA} > .05$$

The operational hypothesis that is represented by the close fit null hypothesis explicitly assumes that the measurement model only approximates the processes that operated in reality to create the observed covariance matrix (Browne & Cudeck, 1993).

Absolute indices of goodness-of-fit directly assess how well a model reproduces the sample data. RMSEA determines the error due to approximation, per degree of freedom of the model (i.e., the discrepancy between $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}(\boldsymbol{\Theta})$ per degree of freedom). The Root Mean Square Error of Approximation is a popular measure of fit that expresses the difference between the observed and estimated sample covariance matrices. According to Diamantopoulos and Siguaw (2000), it is regarded as one of the most informative fit indices as it takes model complexity into consideration. Values below .05 are generally regarded as indicative of good model fit, values above .05 but less than .08 as indicative of reasonable fit, values greater than or equal to .08 but less than .10 are indicative of mediocre fit and values exceeding .10 are generally regarded as indicative of poor fit. Table 5.33 reports a sample RMSEA value of .0515, thus indicating that the measurement model shows good fit in the sample. The p-value for Test of Close Fit (H_{01a}: RMSEA < .05) was 1.00, therefore the close fit null hypothesis H_{01a}: RMSEA ≤ .05 was not rejected (p > .05). The measurement model thus shows close fit. The measurement model is therefore a plausible explanation and the model approximately reproduces the observed covariance matrix, but not perfectly.

The confidence interval for this index was (.0493 ; .0538). Confidence intervals assist in assessing the precision of the fit statistics. The fact that the interval is small indicates a higher level of precision in reflecting the model fit in the population (Byrne, 2001). The fact that the interval includes the critical .05 RMSEA value again provides support for not rejecting the null hypothesis of close fit. If, however, a small RMSEA value with a large confidence interval is indicated, this would suggest that the

estimated discrepancy value is quite imprecise, negating any possibility to accurately determine the degree of fit in the population.

The expected cross-validation index (ECVI) expresses the difference between the reproduced sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ derived from fitting the model on the sample at hand, and the expected covariance matrix that would be obtained in an independent sample of the same size, from the same population (Byrne, 1989; Diamantopoulos & Siguaw, 2000). It therefore focuses on the difference between $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$. Since the model ECVI (8.484) is smaller than the value obtained for the independence model (152.522) but larger than the ECVI value associated with the saturated model (7.712), a model more closely resembling the saturated model seems to have a better chance of being replicated in a cross-validation sample than the fitted model.

The assessment of parsimonious fit acknowledges that model fit can always be improved by adding more paths to the model and estimating more parameters until perfect fit is achieved in the form of a saturated or just-identified model with no degrees of freedom (Kelloway, 1998). In defining and fitting models it would seem essential to find the most parsimonious model that achieves satisfactory fit with as few model parameters as possible (Jöreskog & Sörbom, 1993). The parsimonious normed fit index (PNFI = .896) and the parsimonious goodness-of-fit index (PGFI = .696) approach model fit from this perspective. PNFI and PGFI range from 0 to 1, with higher values indicating a more parsimonious fit. There is no standard for how high either index should be to indicate parsimonious fit (Kelloway, 1998). However, the PNFI was close to reaching the .90 cut-off used for other fit indices. According to Kelloway (1998) these indices are more meaningfully used when comparing two competing theoretical models and are not very useful indicators in this CFA analysis.

The values for this model's Akaike information criterion (AIC= 3893.946) suggest that the fitted measurement model provided a more parsimonious fit than the independent model (70007.784), but not the saturated model (3540.000), since smaller values on these indices indicate a more parsimonious model (Kelloway, 1998). This indicates that the measurement model may lack influential paths. Values for the consistent Akaike information criterion (4684.155) imply that the fitted measurement model provides a more parsimonious fit than both the independent model (70310.526) and the saturated model (12622.271). This provides further support for the fitted model.

Indices of comparative fit that use as a baseline an independence (or null) model, contrast the ability of the model to reproduce the observed covariance matrix with that of a model known *a priori* to fit

the data poorly. The fit indices presented include the normed fit index (NFI= .949), the non-normed fit index (NNFI= .969), the comparative fit index (CFI= .971), the incremental fit index (IFI=0.971) and the relative fit index (RFI =.946).The closer the values are to unity, the better the fit. However, .90 could be considered indicative of a well-fitting model (Diamantopoulos & Siguaw, 2000; Kelloway, 1998). In the current results, all of these indices exceeded the .90 level, which would be indicative of satisfactory comparative fit relative to the independence model.

The critical sample size statistic (CN) refers to the size that the sample would have to reach in order to accept the χ^2 statistic as significant at the .05 significance level (Diamantopoulos & Siguaw, 2000). The estimated CN value (225.152) fell above the recommended threshold value of 200. This threshold is regarded as indicative of the model providing an adequate representation of the data (Diamantopoulos & Siguaw, 2000), although this proposed threshold should be used with caution (Hu & Bentler, 1995).

The standardised RMR may be considered a summary measure of standardised residuals, which represents the average difference between the elements of the sample covariance matrix and the fitted covariance matrix. If the model fit is good, the fitted residuals ($\mathbf{S} - \widehat{\mathbf{\Sigma}}$) should be small in comparison to the magnitude of the elements in **S** (Diamantopoulos & Siguaw, 2000). The RMR (.0803) and standardized RMR (.0616) indicated reasonable fit as values less than .05 on the latter index suggest the model fits the data well (Kelloway, 1998).

The goodness-of-fit index (GFI) and the adjusted goodness-of-fit index (AGFI) reflect how closely the model comes to perfectly reproducing the sample covariance matrix (Diamantopoulos & Siguaw, 2000). The AGFI (.739) adjusts the GFI (.762) for the degrees of freedom in the model (Diamantopoulos & Siguaw, 2000; Jöreskog & Sörbom, 1993) and should be between zero and 1.0 with values exceeding .90 indicating that the model fits the data well (Jöreskog & Sörbom, 1993; Kelloway, 1998). For the fit of this model, both the GFI and AGFI were below the acceptable cut-off level. Kelloway (1998), however, states that GFI and AGFI should be used with some circumspection as guidelines for the interpretation are grounded in experience and therefore somewhat subjective.

In conclusion, when the abovementioned model fit statistics were considered holistically they seem to suggest close fit. In addition, when taking the fitted model, independence model and saturated model into account, evidence was provided in support of the fitted model, however, the model fit may possibly benefit from the inclusion of a few additional paths.

5.7.1.2 MEASUREMENT MODEL RESIDUALS

Residuals refer to the difference between corresponding cells in the observed and fitted covariance/correlation matrices (Diamantopoulos & Siguaw, 2000; Jöreskog & Sörbom, 1993). Standardised residual refers to a residual that is divided by its estimated standard error and can be interpreted as standard normal deviates (i.e., z-scores). It provides valuable diagnostic information on sources of lack of fit in models. Standardised residuals that exceed +2.58 or -2.58 can be considered large (Diamantopoulos & Siguaw, 2000). Residuals should also be distributed approximately symmetrical around zero. Large positive residuals indicate underestimation and imply that additional explanatory paths are needed to rectify the situation. Large negative residuals indicate overestimation and suggest the need to remove some of the paths (Diamantopoulos & Siguaw, 2000; Kelloway, 1998).

The stem-and-leaf plot of the SWSI measurement model for the male sample is depicted in Figure 5.2. This provides graphical information regarding the sample standardised residual distribution. A good model is characterised by a stem-and-leaf plot in which the residuals are distributed approximately symmetrical around zero. An excess of residuals on the positive or negative side would indicate that the covariance terms are systematically under or overestimated.

```
-16|5
 -14|
 -12|
 -10|
 - 8|371
 - 6|84380
 - 4|8876599876665444332000
 - 2|99999887776666655555554444443333222211111100000009999999998888888888777+14
 - 0|99999999999999999999988888888888888888888888777777777777777777777777776+94
   0|11111111111111111111111111111111111111122222222222222222222222222222222+94
   2|00000000000000001111111111111111111122222222222222222222222233333333333+34
   4|000111111222223467888999002234456 99
   6|13
   8|67924
  10|009
```

Figure 5.2. Stem-and-leaf plot of standardised residuals for the SWSI measurement model for the male sample

From the stem-and-leaf plot depicted in Figure 5.2, the distribution of standardised residuals appeared negatively skewed although the median of the distribution was zero. There were, however, a number of both large negative and large positive standardised residuals. The 122 large negative residuals constituted 57% of the total number of unique variance and covariance terms in the observed covariance matrix, and the 153 large positive residuals, 72% (the smallest negative residual was -16.5 and the largest positive residual 10.9). Overall, the prevalence of large positive residuals and fewer large negative residuals would suggest that the observed covariance terms in

the observed covariance matrix are typically underestimated by the derived model parameter estimates. Adding paths to the model may rectify the problem. This confirmed the inference derived earlier from the ECVI and Akaike fit statistics on the possible need of additional paths.

The Q-plot of the SWSI measurement model for the male sample is depicted in Figure 5.3. This is an additional graphical display of residuals by plotting the standardised residuals (horizontal axis) against the quantiles of the normal distribution (Diamantopoulos & Siguaw, 2000). When interpreting the Q-plot, it is important to note the extent to which the data points fall on the 45-degree reference line. Good model fit would be indicated if the points fall on the 45-degree reference line (Jöreskog & Sörbom, 1993). The model fit was less than satisfactory to the extent that the data points swivelled away from the 45-degree reference line. The Q-plot in Figure 5.3 clearly indicates a less than perfect model fit as the standardised residuals tend to deviate from the 45-degree line, especially in the upper and lower regions of the X-axis. This is in line with the results reported in Figure 5.2, where there were both large negative and positive standardised residuals. Subsequently, given the examination of the residuals, it was important to also evaluate the measurement model modification indices.

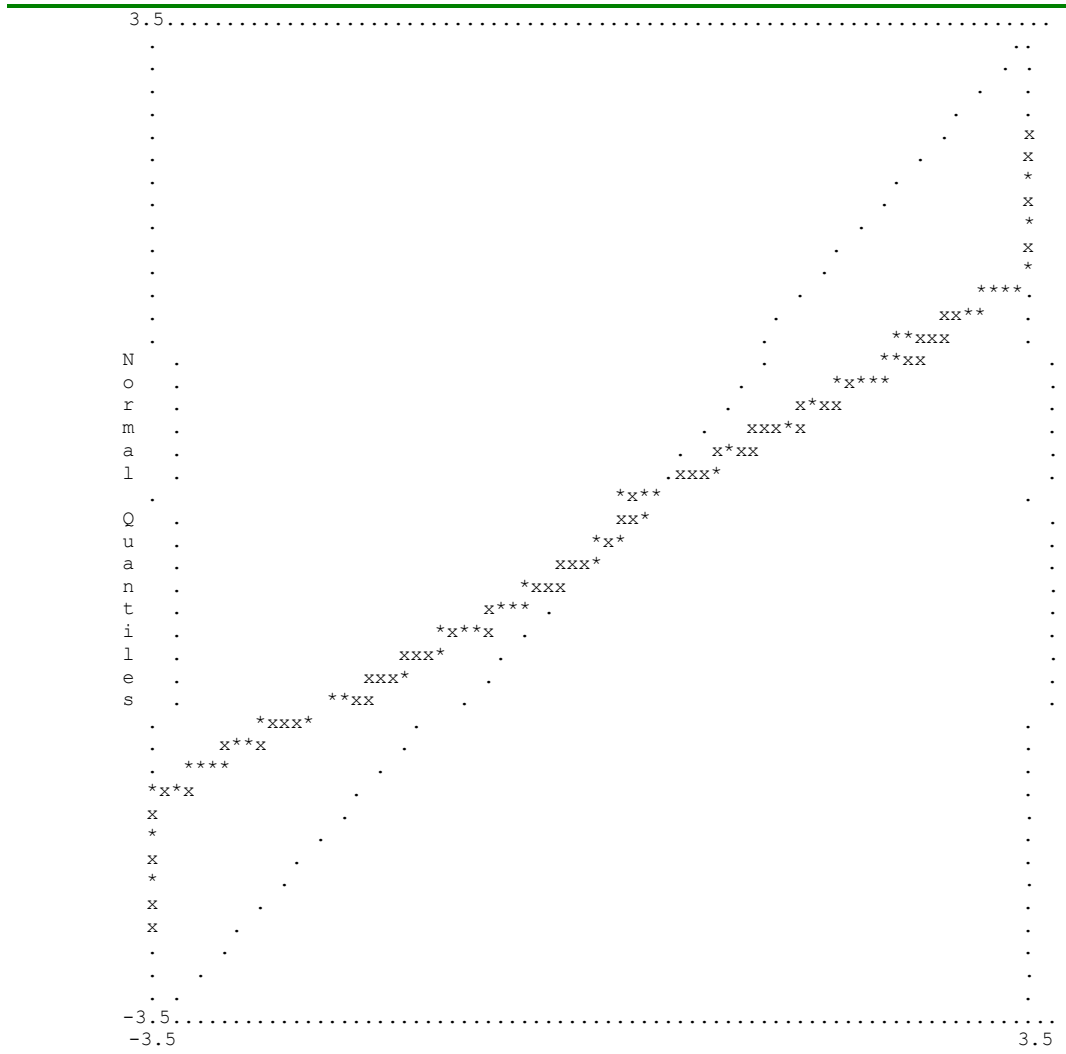

Figure 5.3. Q-plot of SWSI measurement model standardised residuals for the male sample

5.7.1.3 MEASUREMENT MODEL MODIFICATION INDICES

Model modification indices are aimed at answering the question whether any of the currently fixed parameters, when freed in the model, would significantly improve the parsimonious fit of the model. Modification indices (MI) indicate the extent to which the χ^2 fit statistic will decrease if a currently fixed parameter in the model is freed and the model re-estimated (Jöreskog & Sörbom, 1993). Large modification index values (> 6.6349) would be indicative of parameters that, if set free, would improve the fit of the model significantly ($p < .01$; Diamantopoulos & Siguaw, 2000; Jöreskog & Sörbom, 1993). Modifications to the model, based on these statistics, should be theoretically/substantially justified (Diamantopoulos & Siguaw, 2000; Kelloway, 1998). Paths would not be freed in this study as the purpose is to evaluate the fit of the a *priori* model indicated by the test authors. Modification indices calculated for $\boldsymbol{\Lambda_x}$ and $\boldsymbol{\Theta_\delta}$ matrices were, however, examined as

supplementary information on the adequacy of the fitted measurement model. If only a limited number of ways would exist to improve the fit of the model, then this would reflect favourably on the merits of the model derived from the design intentions of the test developers.

Examination of the modification index values calculated for the Λ_x matrix indicated that a number of paths could have been freed that would significantly improve model fit. Seventy-nine of the currently fixed elements in the Λ_x matrix had significant (> 6.6349) modification index values. This suggests that freeing these additional paths would significantly improve the fit of the model[25]. However, the 79 significant modification index values only represent 17% of the possible 472 ways of modifying the factor loading pattern. This small percentage reflects favourably on the fit of the current model.

The magnitude of the predicted factor loadings that should be found if currently fixed elements in the Λ_x matrix were to be freed is reflected in the LISREL output in the completely standardised expected change values. An investigation of the completely standardised expected change matrix showed that none of the loadings were greater than the stringent cut-off value of .71, and therefore freeing a pathway would not result in any potential loadings above .71.

Upon inspection of the theta-delta (Θ_δ) modification indices, approximately 3% (126 out of a possible 3422) of the modification index values in the matrix were significant (> 6.6349). However, upon review of standardised expected changes as a result of freeing off-diagonal error terms in theta-delta, no correlated measurement error terms were proposed. This finding reflects positively on the fit of the measurement model.

However, as previously indicated, no changes were made to the model.

5.7.1.4 INTERPRETATION OF THE SWSI MEASUREMENT MODEL (MALE SAMPLE) PARAMETER ESTIMATES

A measure is designed to provide a valid reflection of a specific latent variable, and therefore the slope of the regression of the observed variables (X_i) on the respective latent variable (ξ_j) in the fitted measurement model has to be substantial and significant (Diamantopoulos & Siguaw, 2000). The unstandardised Λ_x matrix (Table 5.34) contains the regression coefficients of the regression of the manifest variables on the latent variables they were linked to. The regression coefficients of the manifest variables on the latent variables are significant ($p < .05$) if the t-values, as indicated in the

[25] Each of the 79 currently fixed paths would, if freed, result in a significant improvement in model fit. Freeing any of these 79 paths will, however, alter the modification index values in the resultant model.

matrix, exceed |1.96|. Significant indicator loadings provide validity evidence in favour of the indicators (Diamantopoulos & Siguaw, 2000).

Table 5.34

SWSI measurement model unstandardised lambda-x matrix (male sample)

		GEN		RA		REL		TE		CA		JS		LA		WH		WL
			ra1		rel8		te16		ca21		js26		la30		wh37		wl45	
gen1		0.692 (0.042) 16.383*		0.608 (0.046) 13.197*		1.024 (0.044) 23.390*		1.015 (0.041) 24.680*		0.962 (0.051) 18.757*		0.953 (0.048) 20.053*		0.606 (0.051) 11.853*		0.833 (0.049) 17.000*		0.883 (0.046) 19.350*
	ra2		rel9		te17		ca22		js27		la31		wh38		wl46			
gen2		0.740 (0.036) 20.445*		0.751 (0.041) 18.451*		1.002 (0.040) 25.191*		0.959 (0.044) 21.583*		0.912 (0.058) 15.777*		1.021 (0.054) 19.023*		0.734 (0.043) 16.880*		0.628 (0.050) 12.492*		0.859 (0.047) 18.391*
	ra3		rel10		te18		ca23		js28		la32		wh39		wl47			
gen3		0.587 (0.044) 13.269*		0.613 (0.056) 10.890*		1.016 (0.041) 24.899*		1.043 (0.043) 24.255*		1.069 (0.048) 22.130*		0.946 (0.044) 21.653*		0.782 (0.049) 15.830*		0.854 (0.042) 20.247*		0.933 (0.047) 19.745*
	ra4		rel11		te19		ca24		js29		la33		wh40		wl48			
gen4		0.723 (0.043) 16.655*		0.583 (0.046) 12.637*		0.933 (0.040) 23.165*		1.041 (0.039) 26.707*		0.845 (0.050) 17.062*		0.936 (0.041) 22.997*		0.864 (0.045) 19.328*		0.797 (0.042) 18.796*		0.899 (0.040) 22.267*
	ra5		rel12		te20		ca25				la34		wh41		wl49			
gen5		0.653 (0.035) 18.417*		0.665 (0.037) 18.140*		0.877 (0.044) 19.868*		0.836 (0.053) 15.866*		0.896 (0.052) 17.385*				0.863 (0.043) 20.115*		0.664 (0.040) 14.027*		0.986 (0.041) 23.904*
	ra6		rel13								la35		wh42		wl50			
gen6		0.724 (0.032) 22.339*		0.839 (0.048) 17.586*		0.591 (0.045) 13.153*								1.033 (0.046) 22.679*		0.587 (0.042) 14.027*		0.828 (0.046) 17.833*
	ra7		rel14								la36		wh43					
gen7		0.732 (0.046) 15.768*		0.773 (0.037) 21.062*		0.837 (0.044) 19.029*								0.798 (0.046) 17.483*		0.389 (0.034) 11.438*		
			rel15										wh44					
gen8		0.679 (0.035) 19.262*				0.726 (0.042) 17.260*										0.758 (0.050) 15.216*		
gen9		0.602 (0.046) 12.996*																

*t-values > |1.96| indicates significant path coefficients; values in brackets represent standard error estimates

In Table 5.34, the intercepts reflect the average item score when the latent stress dimension is zero. These estimates should not be interpreted literally but rather as necessary estimates along with the slope and error variance estimates to describe the nature of the regression of X_i on ξ_j in each gender group, and eventually to compare the nature of the regression of X_i on ξ_j across gender groups.

Table 5.34 indicates that 49 (83%) of the lambda estimates were statistically significantly different from zero ($p < .05$). Furthermore, the regression coefficients of the manifest variables on the latent variables are significant ($p < .05$), as the *t*-values exceed $|1.96|$. Therefore, H_{07} to H_{065} can all be rejected in favour of H_{a7} to H_{a65}. These significant indicator loadings provided validity evidence in favour of the indicators (Diamantopoulos & Siguaw, 2000). By removing any of these pathways, model fit would deteriorate significantly (Kelloway, 1998).

However, Diamantopoulos and Siguaw (2000) warn against solely relying on unstandardised loadings as it might be hard to compare the validity of different indicators measuring a particular construct. Therefore, the completely standardised factor loading matrix ($\mathbf{\Lambda_x}$) was investigated. The completely standardised estimates indicate the average change in standard deviation units in the observed variable X directly resulting from a one standard deviation change in a latent variable $\boldsymbol{\xi}$ to which it has been linked, holding the effect of all other variables constant. The completely standardised factor loading matrix is presented in Table 5.35. The completely standardised factor loadings can be interpreted as correlation coefficients. By using a stringent cut-off point of .71, thirty-three of the 59 factor loadings appeared satisfactory. The 26 items that returned factor loadings less than .71 were of the most concern.

Table 5.35

SWSI measurement model completely standardised solution lambda-x matrix (male sample)

gen1	**gen2**	**gen3**	**gen4**	**gen5**	**gen6**	**gen7**	**gen8**	**gen9**	**ra1**
.703	.767	.608	.666	.727	.784	.653	.753	.597	.597
ra2	**ra3**	**ra4**	**ra5**	**ra6**	**ra7**	**rel8**	**rel9**	**rel10**	**rel11**
.697	.502	.555	.720	.697	.721	.795	.821	.824	.800
rel12	**rel13**	**rel14**	**rel15**	**te16**	**te17**	**te18**	**te19**	**te20**	**ca21**
.772	.606	.753	.722	.824	.772	.842	.871	.665	.712
ca22	**ca23**	**ca24**	**ca25**	**js26**	**js27**	**js28**	**js29**	**la30**	**la31**
.623	.791	.672	.703	.763	.784	.797	.811	.539	.658
la32	**la33**	**la34**	**la35**	**la36**	**wh37**	**wh38**	**wh39**	**wh40**	**wh41**
.658	.725	.701	.774	.662	.666	.531	.725	.716	.696
wh42	**wh43**	**wh44**	**wl45**	**wl46**	**wl47**	**wl48**	**wl49**	**wl50**	
.600	.554	.616	.719	.727	.747	.783	.799	.729	

Total variance in the i^{th} individual item ($\mathbf{X_i}$) consists of (a) variance in the latent variable the individual item was designed to reflect ($\boldsymbol{\xi_j}$), (b) variance in other systematic latent effects the individual item was not designed to reflect, as well as (c) random measurement error. The

measurement error term ($\boldsymbol{\delta_i}$), in the model specification, accounts for the latter two sources of variance in the individual item. The square of the completely standardised factor loadings given in Table 5.35 indicates the proportion of indicator variance explained in terms of the latent variable it is meant to express (Diamantopoulos and Siguaw, 2000).

Since each indicator variable only loads on a single latent variable, the squared completely standardised loadings equal the R^2 values shown in Table 5.36. The values shown in Table 5.36 could therefore be interpreted as indicator variable validity coefficients, $\rho(Xi, \xi j)$. Since ($\lambda_{ij}^2 + \theta_{\delta ii}$) are equal to unity in the completely standardised solution, the validity coefficients, $\rho(X_i, \xi_j)$ can be defined as follows in Equation 3:

$\rho(X_i, \xi_j)$ = σ^2systematic-relevant/(σ^2systematic-relevant + σ^2non-relevant)

$= \lambda_{ij^2}/[\lambda_{ij^2} + \Theta\delta_{ii}]$

$= 1 - (\Theta\delta_i/[\lambda_{ij^2} + \Theta\delta_{ii}])$

$= 1 - \Theta\delta_{ii}$

$= \lambda_{ij^2}$----------(3)

Since reliability could be defined as the extent to which variance in indicator variables can be attributed to systematic sources, irrespective of whether the source of variance is relevant to the measurement intention or not, the values shown in Table 5.36 could simultaneously be interpreted as lower bound estimates of the item reliabilities (Diamantopoulos & Siguaw, 2000; Jöreskog & Sörbom, 1996a). The extent to which the true item reliabilities would be under-estimated would be determined by the extent to which $\boldsymbol{\delta_{ii}}$ contains the effect of the systematic non-relevant latent influences. A high R^2 value would indicate that variance in the indicator in question, to a large degree, reflects variance in the latent variable to which it has been linked. The rest of the variance, which is not explained by the latent variable, can be ascribed to systematic and random measurement error (Diamantopoulos & Siguaw, 2000).

In terms of the foregoing argument, the values of the squared multiple correlations (R^2) for the indicator variables shown in Table 5.36 indicate that the items provided relatively contaminated reflections of their designed dimension. Thirty-three items (55.9%) reported satisfactory R^2 values exceeding .50 (λ_{ij}^2=.71^2), indicating that 50% or more of the variance in the item can be explained by the latent variable the item is meant to reflect (i.e., Tools and Equipment). Approximately 44% of the individual indicators explained less than half of the variance in the latent variable they were meant to reflect.

Table 5.36

SWSI measurement model squared multiple correlations for X-variables (male sample)

gen1	gen2	gen3	gen4	gen5	gen6	gen7	gen8	gen9	ra1
.494	.588	.370	.444	.529	.615	.426	.567	.357	.486
ra2	**ra3**	**ra4**	**ra5**	**ra6**	**ra7**	**rel8**	**rel9**	**rel10**	**rel11**
.486	.252	.308	.519	.485	.520	.632	.674	.679	.640
rel12	**rel13**	**rel14**	**rel15**	**te16**	**te17**	**te18**	**te19**	**te20**	**ca21**
.596	.368	.566	.521	.679	.596	.710	.758	.442	.506
ca22	**ca23**	**ca24**	**ca25**	**js26**	**js27**	**js28**	**js29**	**la30**	**la31**
.389	.625	.452	.495	.582	.615	.636	.658	.291	.434
la32	**la33**	**la34**	**la35**	**la36**	**wh37**	**wh38**	**wh39**	**wh40**	**wh41**
.433	.525	.491	.599	.438	.443	.282	.526	.513	.484
wh42	**wh43**	**wh44**	**wl45**	**wl46**	**wl47**	**wl48**	**wl49**	**wl50**	
.360	.307	.379	.518	.528	.558	.613	.638	.531	

Table 5.37 contains the completely standardised error variance of the i^{th} indicator variable ($\Theta_{\delta ii}$), which consists of the systematic non-relevant variance and random error variance (refer to Equation 3).

Table 5.37

SWSI measurement model completely standardised theta-delta matrix (male sample)

gen1	gen2	gen3	gen4	gen5	gen6	gen7	gen8	gen9	ra1
.506	.412	.630	.556	.471	.385	.574	.433	.644	.643
ra2	**ra3**	**ra4**	**ra5**	**ra6**	**ra7**	**rel8**	**rel9**	**rel10**	**rel11**
.514	.748	.692	.481	.515	.480	.368	.326	.321	.360
rel12	**rel13**	**rel14**	**rel15**	**te16**	**te17**	**te18**	**te19**	**te20**	**ca21**
.404	.632	.434	.479	.321	.404	.290	.242	.558	.494
ca22	**ca23**	**ca24**	**ca25**	**js26**	**js27**	**js28**	**js29**	**la30**	**la31**
.611	.375	.548	.505	.418	.385	.364	.342	.709	.566
la32	**la33**	**la34**	**la35**	**la36**	**wh37**	**wh38**	**wh39**	**wh40**	**wh41**
.567	.475	.509	.401	.562	.557	.718	.474	.487	.516
wh42	**wh43**	**wh44**	**wl45**	**wl46**	**wl47**	**wl48**	**wl49**	**wl50**	
.640	.693	.621	.482	.472	.442	.387	.362	.469	

The phi-matrix of correlations between the 9 latent occupational stress sub-scales is provided in Table 5.38. Twenty-one (58%) of the correlations in Table 5.38 are statistically significant ($p < .05$). The off-diagonal elements of the Φ matrix are the occupational stress sub-scale correlations

corrected for the attenuating effect of measurement error. As the **Φ** matrix is a positive definite, and off-diagonal entries tend to contain relatively moderate correlations, the results tend to provide some support for discriminant validity of the occupational stress sub-scales.

Table 5.38

SWSI measurement model completely standardised phi matrix (male sample)

	GEN	RA	REL	TE	CA	JS	LA	WH	WL
GEN	1.000								
RA	0.594	1.000							
REL	0.298	0.626	1.000						
TE	0.250	0.482	0.569	1.000					
CA	0.260	0.478	0.570	0.445	1.000				
JS	0.424	0.569	0.509	0.410	0.690	1.000			
LA	0.402	0.653	0.778	0.624	0.669	0.694	1.000		
WH	0.600	0.562	0.368	0.311	0.359	0.462	0.456	1.000	
WL	0.569	0.478	0.221	0.258	0.230	0.329	0.357	0.717	1.000

5.7.2 SINGLE-GROUP MEASUREMENT MODEL FIT: FEMALE SAMPLE

Operational hypothesis 1b was tested by testing H_{01b_female}: RMSEA ≤ .05. The fit of the occupational stress measurement model for the female sample was evaluated based on the array of model fit indices reported by LISREL. The standardised residuals, modification indices, as well as the measurement model parameter estimates were subsequently evaluated.

A visual representation of the SWISI measurement model that was fitted to the data of the female sample (N=460) is shown in Figure 5.4.

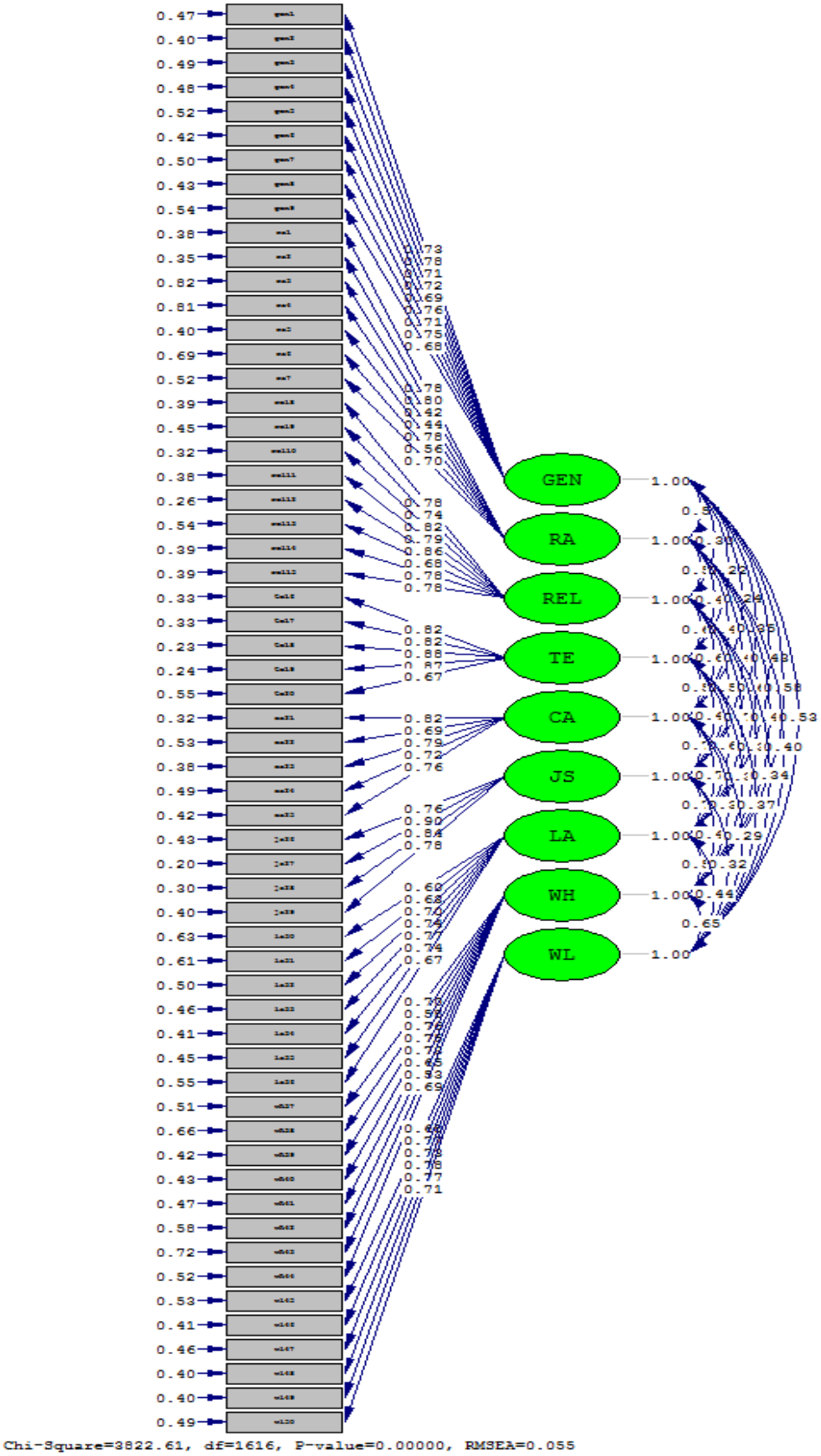

Figure 5.4. Representation of the fitted SWSI measurement model: female sample

5.7.2.1 MEASUREMENT MODEL FIT INDICES

The measurement model converged in 21 iterations. The spectrum of fit statistics is shown in Table 5.39.

Table 5.39
Goodness of fit statistics for the SWSI measurement model: female sample

Degrees of Freedom = 1616
Minimum Fit Function Chi-Square = 4316.100 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 4673.276 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 3822.614 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 2206.614
90 Percent Confidence Interval for NCP = (2029.745 ; 2391.101)

Minimum Fit Function Value = 9.403
Population Discrepancy Function Value (F0) = 4.807
90 Percent Confidence Interval for F0 = (4.422 ; 5.209)
Root Mean Square Error of Approximation (RMSEA) = 0.0545
90 Percent Confidence Interval for RMSEA = (0.0523 ; 0.0568)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000

Expected Cross-Validation Index (ECVI) = 8.999
90 Percent Confidence Interval for ECVI = (8.614 ; 9.401)
ECVI for Saturated Model = 7.712
ECVI for Independence Model = 170.466

Chi-Square for Independence Model with 1711 Degrees of Freedom = 78125.692
Independence AIC = 78243.692
Model AIC = 4130.614
Saturated AIC = 3540.000
Independence CAIC = 78546.434
Model CAIC = 4920.823
Saturated CAIC = 12622.271

Normed Fit Index (NFI) = 0.951
Non-Normed Fit Index (NNFI) = 0.969
Parsimony Normed Fit Index (PNFI) = 0.898
Comparative Fit Index (CFI) = 0.971
Incremental Fit Index (IFI) = 0.971
Relative Fit Index (RFI) = 0.948

Critical N (CN) = 211.274

Root Mean Square Residual (RMR) = 0.0881
Standardized RMR = 0.0622
Goodness of Fit Index (GFI) = 0.743
Adjusted Goodness of Fit Index (AGFI) = 0.719
Parsimony Goodness of Fit Index (PGFI) = 0.679

Upon first inspection of Table 5.39, the degrees of freedom were indicated to be 1616. This corresponds with the earlier calculations (section 4.5.1.2) and provided evidence that the model was specified correctly.

As with the male sample, the exact fit null hypothesis was not tested.

The following close fit null hypothesis was tested:

$$H_{01b\_female}: RMSEA \leq .05$$

$$H_{a1b\_female}: RMSEA > .05$$

A RMSEA value of .0545 was obtained (Table 5.39), indicating that the measurement model showed good fit. The p-value for Test of Close Fit (H_{01b}: RMSEA < .05) was 1.00, and therefore the close fit null hypothesis, H_{01b}: RMSEA ≤ .05, was not rejected (p > .05). The measurement model therefore showed close good fit. Based on these results it may be concluded that the measurement model was therefore a plausible explanation and the model approximately reproduced the observed covariance matrix, but not perfectly.

The confidence interval for this index was (.0523 ; .0568). The fact that the interval was small indicated a higher level of precision in reflecting the model fit in the population (Byrne, 2001).

Since the model ECVI (8.999) was smaller than the value obtained for the independence model (170.466) but larger than the ECVI value associated with the saturated model (7.712), a model more closely resembling the saturated model seems to have a better chance of being replicated in a cross-validation sample than the fitted model.

Parsimonious normed fit index (PNFI) and parsimonious goodness-of-fit index (PGFI) range from 0 to 1, with higher values indicating a more parsimonious fit. The PNFI for this model was .898 which is close to reaching the .90 cut-off used for other fit indices. However, the PGFI was slightly lower at .679. According to Kelloway (1998) these indices are more meaningfully used when comparing two competing theoretical models and are not very useful indicators in this CFA analysis.

The values for this model's Akaike information criterion (AIC = 4130.614) suggest that the fitted measurement model provided a more parsimonious fit than the independent model (78243.692) but not the saturated model (3540.000) since smaller values on these indices indicate a more parsimonious model (Kelloway, 1998). This indicates that the measurement model may lack influential paths. Values for the consistent Akaike information criterion (4920.823) imply that the fitted measurement model provides a more parsimonious fit than both the independent model (78546.434) and the saturated model (12622.271). This provided further support for the fitted model.

The comparative fit indices returned the following values: normed fit index (NFI = .951), the non-normed fit index (NNFI = .969), the comparative fit index (CFI = .971), the incremental fit index (IFI =

.971) and the relative fit index (RFI = .948). In the current results, all of these indices exceed the .90 level, which would be indicative of satisfactory comparative fit relative to the independence model.

The estimated CN value (211.274) fell above the recommended threshold value of 200. This threshold is regarded as indicative of the model providing an adequate representation of the data (Diamantopoulos & Siguaw, 2000) although this proposed threshold should be interpreted with caution (Hu & Bentler, 1995).

The RMR (.0881) and standardized RMR (.0622) indicated reasonable fit as values less than .05 on the latter index suggest the model fits the data well (Kelloway, 1998).

For the fit of this model, both the GFI (.743) and AGFI (.719) are below the acceptable cut-off level of .90, reflecting that the model does not perfectly reproduce the same covariance matrix (Diamantopoulos & Siguaw, 2000). For the fit of this model, both the GFI and AGFI are below the acceptable cut-off level.

In conclusion, when the abovementioned model fit statistics were considered holistically they seem to suggest close fit. In addition, when taking the fitted model, independence model and saturated model into account, evidence was provided in support of the fitted model, however, the model fit may possibly benefit from the inclusion of a few additional paths.

5.7.2.2 MEASUREMENT MODEL RESIDUALS

The stem-and-leaf plot of the SWSI measurement model for the female sample is depicted in Figure 5.5. This provides graphical information regarding the sample standardised residual distribution.

```
- 3|1
 - 2|
 - 2|
 - 1|
 - 1|320
 - 0|98877776666666555555555
 - 0|4444444444444444444444443333333333333333333333333333333333333333333333333333+97
   0|1111111111111111111111111111111111111111111111111111111111111111111111111111+93
   0|55555555555555555666666666777888889
   1|01
   1|
   2|1
```

Figure 5.5. Stem-and-leaf plot of standardised residuals for the SWSI measurement model for the female sample

From the stem-and-leaf plot depicted in Figure 5.5, the distribution of standardised residuals appears essentially symmetrical with the median of the distribution at zero. The largest positive standardised residual was 20.914, however the smallest negative residual was -31.275. The 137 large

positive standardised residuals constituted 64% of the total number of unique variance and covariance terms in the observed covariance matrix, while the 111 large negative standardised residuals constituted 52%. Overall, the prevalence of positive residuals would suggest that the observed covariance terms in the observed covariance matrix are typically underestimated by the derived model parameter estimates. Adding paths to the model may rectify the problem.

The Q-plot of the SWSI measurement model for the female sample is depicted in Figure 5.6. Based upon the extent to which the data points fall on the 45-degree reference line, the Q-plot in Figure 5.6 clearly indicates a less than perfect model fit as the standardised residuals tend to deviate from the 45-degree line, especially in the upper and lower regions of the X-axis. Given the examination of the residuals, it was important to also evaluate the measurement model modification indices.

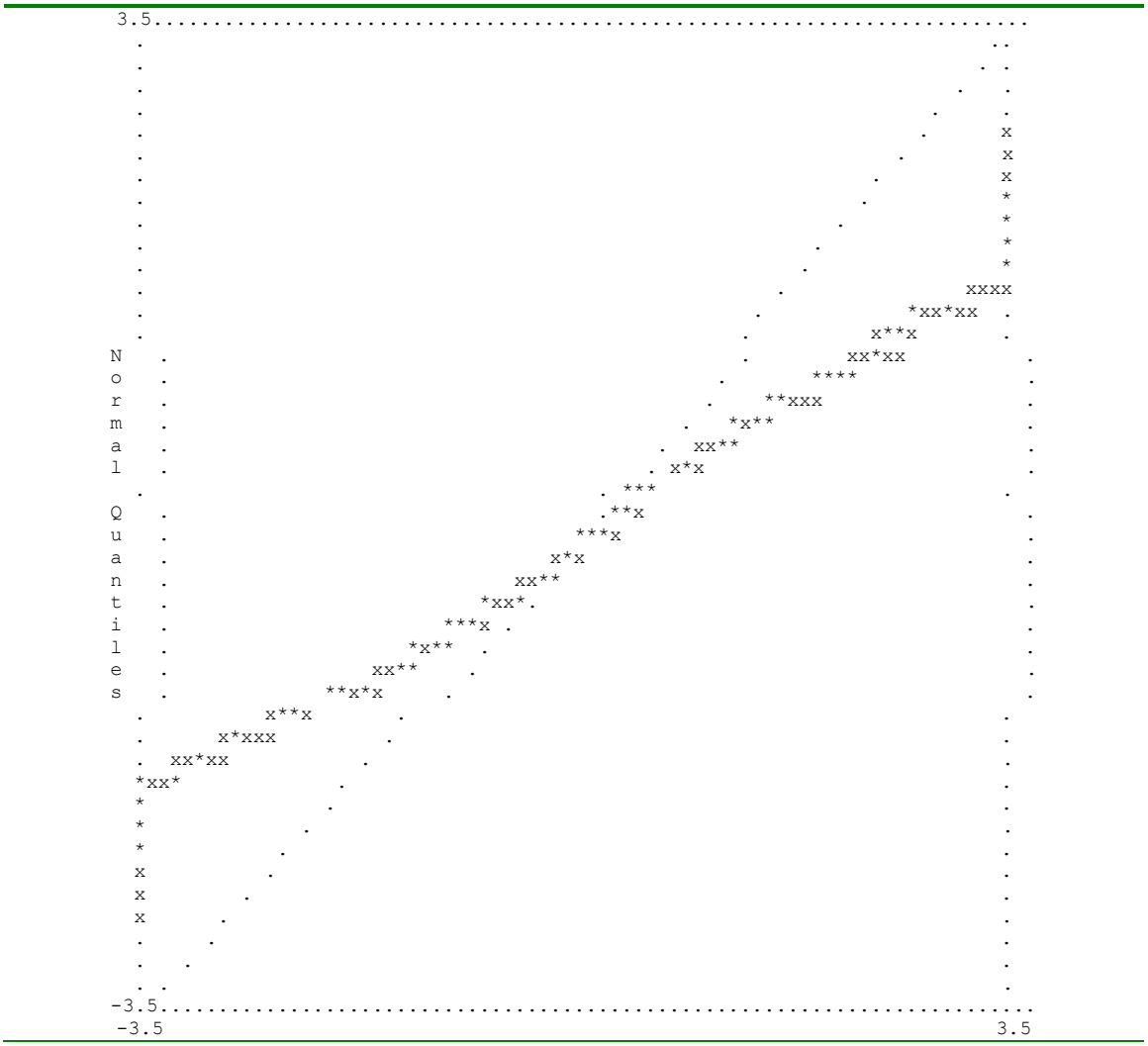

Figure 5.6. Q-plot of the SWSI measurement model standardised residuals for the female sample

5.7.2.3 MEASUREMENT MODEL MODIFICATION INDICES

Examination of the modification index values calculated for the $\boldsymbol{\Lambda_x}$ matrix indicated that a number of paths could be freed that would significantly improve model fit. Approximately 16% of the currently fixed elements in the $\boldsymbol{\Lambda_x}$ matrix were identified as being significant (> 6.6349). This suggested that freeing these additional paths would significantly improve the fit of the model. However, the 75 significant modification index values only represented 16% of the possible 472 ways of modifying the factor loading pattern. This small percentage reflects favourable on the fit of the current model.

An investigation of the completely standardised expected change matrix showed that none of the loadings were greater than the stringent cut-off value of .71, and therefore freeing a particular pathway would not result in a corresponding potential loading above .71.

Upon inspection of the theta-delta ($\boldsymbol{\Theta_\delta}$) modification indices, only approximately 5% (159 out of a possible 3422) of the modification index values in the matrix were significant (> 6.6349). However, upon review of standardised expected changes as a result of possibly freeing off-diagonal error terms in theta-delta, no correlated measurement error terms could be proposed. This finding reflects positively on the fit of the measurement model.

Due to the nature of the study no changes were made to the model.

5.7.2.4 INTERPRETATION OF THE SWSI MEASUREMENT MODEL (FEMALE SAMPLE) PARAMETER ESTIMATES

The unstandardised $\boldsymbol{\Lambda_x}$ matrix (Table 5.40) contains the regression coefficients of the regression of the manifest variables on the latent variables they were linked to. The regression coefficients of the manifest variables on the latent variables are significant ($p < .05$) if the *t*-values, as indicated in the matrix, exceed |1.96|. Significant indicator loadings provide validity evidence in favour of the indicators (Diamantopoulos & Siguaw, 2000).

Table 5.40

SWSI measurement model unstandardised lambda-X matrix (female sample)

	GEN		RA		REL		TE		CA		JS		LA		WH		WL
gen1	0.711 (0.041) 17.136*	ra1	0.897 (0.043) 20.828*	rel8	1.045 (0.044) 23.712*	te16	0.976 (0.044) 21.982*	ca21	1.131 (0.046) 24.361*	js26	1.010 (0.056) 18.162*	la30	0.666 (0.050) 13.455*	wh37	0.901 (0.051) 17.732*	wl45	0.868 (0.046) 18.795*
gen2	0.827 (0.041) 19.988*	ra2	0.915 (0.041) 22.308*	rel9	0.944 (0.051) 18.644*	te17	1.022 (0.044) 23.053*	ca22	0.971 (0.056) 17.386*	js27	1.216 (0.042) 28.797*	la31	0.744 (0.052) 14.339*	wh38	0.760 (0.054) 14.068*	wl46	0.933 (0.049) 19.118*
gen3	0.756 (0.047) 16.017*	ra3	0.503 (0.053) 9.403*	rel10	1.084 (0.044) 24.830*	te18	1.118 (0.039) 28.580*	ca23	1.087 (0.055) 19.610*	js28	1.045 (0.042) 24.687*	la32	0.841 (0.051) 16.577*	wh39	0.987 (0.048) 20.527*	wl47	0.934 (0.050) 18.845*
gen4	0.817 (0.045) 18.114*	ra4	0.456 (0.049) 9.326*	rel11	0.962 (0.049) 19.538*	te19	1.096 (0.043) 25.755*	ca24	0.915 (0.051) 17.821*	js29	0.934 (0.047) 19.968*	la33	0.908 (0.047) 19.282*	wh40	0.931 (0.044) 21.046*	wl48	0.898 (0.043) 20.683*
gen5	0.684 (0.041) 16.623*	ra5	0.813 (0.042) 19.425*	rel12	1.068 (0.045) 23.748*	te20	0.907 (0.053) 16.995*	ca25	1.057 (0.049) 21.545*			la34	0.999 (0.048) 20.835*	wh41	0.819 (0.046) 17.919*	wl49	1.007 (0.050) 20.117*
gen6	0.694 (0.038) 18.144*	ra6	0.646 (0.050) 12.834*	rel13	0.752 (0.048) 15.685*							la35	1.214 (0.065) 18.764*	wh42	0.716 (0.052) 13.783*	wl50	0.845 (0.049) 17.426*
gen7	0.769 (0.045) 17.142*	ra7	0.790 (0.046) 17.332*	rel14	0.971 (0.047) 20.806*							la36	0.865 (0.050) 17.130*	wh43	0.447 (0.043) 10.477*		
gen8	0.744 (0.041) 18.086*			rel15	0.932 (0.051) 18.449*									wh44	0.916 (0.052) 17.648*		
gen9	0.702 (0.045) 15.565*																

*t-values > |1.96| indicates significant path coefficients; values in brackets represent standard error estimates

In Table 5.40, the intercepts reflect the average item score when the latent stress dimension is zero. These estimates should, however, not be interpreted literally but rather as necessary estimates along with the slope and error variance estimates to describe the nature of the regression of X_i on ξ_j in each gender group and eventually to compare the nature of the regression of X_i on ξ_j across gender groups. Table 5.40 indicates that 39 (66%) of the lambda estimates were statistically significantly different from zero ($p < .05$). Furthermore, the regression coefficients of the manifest variables on the latent variables are significant ($p < .05$) as the t-values exceed |1.96|. Therefore, H_{07}

to H_{065} can all be rejected in favour of H_{a7} to H_{a65}. These significant indicator loadings provide validity evidence in favour of the indicators (Diamantopoulos & Siguaw, 2000). By removing any of these pathways, model fit would deteriorate significantly (Kelloway, 1998).

However, Diamantopoulos and Siguaw (2000) warn against solely relying on unstandardized loadings as it might be hard to compare the validity of different indicators measuring a particular construct. Therefore the completely standardised factor loading matrix ($\mathbf{\Lambda_x}$) was investigated. The completely standardised factor loading matrix is presented in Table 5.41. The completely standardised factor loadings can be interpreted as correlation coefficients. By using a stringent cut-off point of .71, 40 of the 59 factor loadings appeared satisfactory. The 19 items that returned factor loadings less than .71 were of the most concern.

Table 5.41

SWSI measurement model completely standardised solution lambda-x matrix (female sample)

gen1	**gen2**	**gen3**	**gen4**	**gen5**	**gen6**	**gen7**	**gen8**	**gen9**	**ra1**
.727	.776	.712	.718	.693	.761	.707	.752	.675	.784
ra2	**ra3**	**ra4**	**ra5**	**ra6**	**ra7**	**rel8**	**rel9**	**rel10**	**rel11**
.804	.421	.436	.776	.555	.696	.779	.742	.822	.790
rel12	**rel13**	**rel14**	**rel15**	**te16**	**te17**	**te18**	**te19**	**te20**	**ca21**
.863	.680	.783	.779	.816	.816	.877	.872	.672	.822
ca22	**ca23**	**ca24**	**ca25**	**js26**	**js27**	**js28**	**js29**	**la30**	**la31**
.686	.785	.717	.759	.758	.897	.838	.776	.604	.628
la32	**la33**	**la34**	**la35**	**la36**	**wh37**	**wh38**	**wh39**	**wh40**	**wh41**
.705	.737	.771	.741	.674	.697	.582	.764	.754	.728
wh42	**wh43**	**wh44**	**wl45**	**wl46**	**wl47**	**wl48**	**wl49**	**wl50**	
.647	.532	.691	.683	.765	.735	.777	.775	.713	

The square of the completely standardised factor loadings given in Table 5.42 indicates the proportion of indicator variance explained in terms of the latent variable it is meant to express (Diamantopoulos and Siguaw, 2000). Since each indicator variable only loads on a single latent variable the squared completely standardised loadings equal the R^2 values shown in Table 5.42. The squared multiple correlations (R^2) indicate that the items provided relatively contaminated reflections of their designed dimension.

Forty items (67.7%) obtained satisfactory values equal to or greater than .50, indicating that 50% or more of the variance in the item can be explained by the latent variable the item was meant to

reflect. The rest of the variance, which is not explained by the latent variable, could be ascribed to systematic and random measurement error (Diamantopoulos and Siguaw, 2000). Approximately 32% of the items had less than half of their variance explained by the latent variable they were meant to reflect.

Table 5.42

SWSI measurement model squared multiple squared correlations for X-variable (female sample)

gen1	gen2	gen3	gen4	gen5	gen6	gen7	gen8	gen9	ra1
.529	.602	.507	.516	.480	.579	.500	.565	.456	.615
ra2	**ra3**	**ra4**	**ra5**	**ra6**	**ra7**	**rel8**	**rel9**	**rel10**	**rel11**
.646	.177	.190	.602	.308	.484	.606	.550	.675	.624
rel12	**rel13**	**rel14**	**rel15**	**te16**	**te17**	**te18**	**te19**	**te20**	**ca21**
.745	.463	.614	.607	.666	.666	.769	.760	.451	.676
ca22	**ca23**	**ca24**	**ca25**	**js26**	**js27**	**js28**	**js29**	**la30**	**la31**
.470	.617	.514	.577	.574	.804	.702	.602	.365	.395
la32	**la33**	**la34**	**la35**	**la36**	**wh37**	**wh38**	**wh39**	**wh40**	**wh41**
.497	.544	.594	.550	.454	.485	.339	.584	.569	.530
wh42	**wh43**	**wh44**	**wl45**	**wl46**	**wl47**	**wl48**	**wl49**	**wl50**	
.419	.283	.478	.466	.585	.540	.603	.600	.509	

Table 5.43 indicates the completely standardised error variance of the i^{th} indicator variable ($\Theta_{\delta ii}$), which consists of the systematic non-relevant variance and random error variance (refer to Equation 3).

Table 5.43

SWSI measurement model completely standardised theta-delta matrix (female sample)

gen1	gen2	gen3	gen4	gen5	gen6	gen7	gen8	gen9	ra1
.471	.398	.493	.484	.520	.421	.500	.435	.544	.385
ra2	**ra3**	**ra4**	**ra5**	**ra6**	**ra7**	**rel8**	**rel9**	**rel10**	**rel11**
.354	.823	.810	.398	.692	.516	.394	.450	.325	.376
rel12	**rel13**	**rel14**	**rel15**	**te16**	**te17**	**te18**	**te19**	**te20**	**ca21**
.255	.537	.386	.393	.334	.334	.231	.240	.549	.324
ca22	**ca23**	**ca24**	**ca25**	**js26**	**js27**	**js28**	**js29**	**la30**	**la31**
.530	.383	.486	.423	.426	.196	.298	.398	.635	.605
la32	**la33**	**la34**	**la35**	**la36**	**wh37**	**wh38**	**wh39**	**wh40**	**wh41**
.503	.456	.406	.450	.546	.515	.661	.416	.431	.470
wh42	**wh43**	**wh44**	**wl45**	**wl46**	**wl47**	**wl48**	**wl49**	**wl50**	
.581	.717	.522	.534	.415	.460	.397	.400	.491	

The phi-matrix of correlations between the 9 latent occupational stress sub-scales is provided in Table 5.44. Twenty-four (67%) of the correlations in Table 5.44 were statistically significant. The off-diagonal elements of the **Φ** matrix are the occupational stress scale correlations corrected for the attenuating effect of measurement error. As the **Φ** matrix is a positive definite, and off-diagonal entries tend to contain relatively moderate correlations, the results tend to provide some support for discriminant validity of the occupational stress sub-scales.

Table 5.44

SWSI measurement model completely standardised phi matrix (female sample)

	GEN	RA	REL	TE	CA	JS	LA	WH	WL
GEN	1.000								
RA	0.572	1.000							
REL	0.304	0.561	1.000						
TE	0.218	0.471	0.600	1.000					
CA	0.239	0.435	0.642	0.567	1.000				
JS	0.349	0.477	0.508	0.419	0.724	1.000			
LA	0.430	0.597	0.764	0.641	0.729	0.699	1.000		
WH	0.581	0.416	0.308	0.334	0.326	0.409	0.497	1.000	
WL	0.527	0.402	0.335	0.370	0.286	0.322	0.442	0.648	1.000

5.7.3 SUMMARY OF SWSI SINGLE-GROUP MEASUREMENT MODEL FIT

It could be concluded that the SWSI single-group measurement model fitted well for both gender samples independently, as this was supported by the array of indices inspected. These results, to a certain degree, suggested that the SWSI instrument measures the same latent variable in both gender groups, suggesting the possibility of lack of construct bias. However, a more formal, rigorous test of lack of construct bias was required.

5.8 DISCRIMINANT VALIDITY

Following the fitting of the single-group measurement model for each gender sample, the discriminant validity of the SWSI measure was evaluated for each gender sample. Discriminant validity was evaluated in order to investigate the extent to which each construct (represented by a different sub-scale) in the instrument may be considered to be truly distinct from the other constructs in the SWSI, given the manner in which the constructs are measured. As described in Chapter 4, two discriminant validity tests were performed in order to determine whether the constructs are unique and capture some phenomena the other measures do not capture. The 9 sub-scales were subjected to the ‘average variance extracted’ versus ‘shared variance’ tests, as well as

the calculation of 95% confidence interval estimates for the 36 correlations between the latent stress dimensions for each gender sample.

5.8.1 AVERAGE VARIANCE EXTRACTED VERSUS SHARED VARIANCE: MALE SAMPLE

One of the ways in which discriminant validity is assessed is by comparing the average variance extracted proportions for any two constructs with the square of the correlation estimate between these two constructs (shared variance). In order to conclude that a particular latent stress construct (i.e. sub-scale) explains its item measures better than it explains another stress construct (i.e. sub-scale), the variance extracted estimates should be greater than the squared correlation estimate. The average variance extracted should also exceed at least .50 so that the latent variable being measured by the indicators account for a larger proportion of the variance in the indicators than measurement error. The results are displayed in Table 5.45.

Table 5.45

Average variance extracted versus squared correlation: male sample

SCALE GEN	ESTIMATE	SQUARED CORRELATION	AVERAGE FOR GEN	AVERAGE FOR SCALE
RA	.594	.352836	.487651296	.418084416
REL	.298	.088804	.487651296	.584528823
TE	.250	.062500	.487651296	.636984753
CA	.260	.067600	.487651296	.493354098
JS	.424	.179776	.487651296	.622632544
LA	.402	.161604	.487651296	.458698433
WH	.600	.360000	.487651296	.411793086
WL	.569	.323761	.487651296	.564364554

SCALE RA	ESTIMATE	SQUARED CORRELATION	AVERAGE FOR RA	AVERAGE FOR SCALE
REL	.626	.391876	.418084416	.584528823
TE	.482	.232324	.418084416	.636984753
CA	.478	.228484	.418084416	.494657696
JS	.569	.323761	.418084416	.622632544

LA	.653	.426409	**.418084416**	.458698433
WH	.562	.315844	.418084416	.411793086
WL	.478	.228484	.418084416	.564364554

SCALE REL	ESTIMATE	SQUARED CORRELATION	AVERAGE FOR REL	AVERAGE FOR SCALE
TE	.569	.323761	.584528823	.636984753
CA	.570	.324900	.584528823	.493354098
JS	.509	.259081	.584528823	.622632544
LA	.778	.605284	**.584528823**	**.460756862**
WH	.368	.135424	.584528823	.411793086
WL	.221	.048841	.584528823	.564364554

SCALE TE	ESTIMATE	SQUARED CORRELATION	AVERAGE FOR TE	AVERAGE FOR SCALE
CA	.445	.198025	.636984753	.493354098
JS	.410	.168100	.636984753	.622632544
LA	.624	.389376	.636984753	.458698433
WH	.311	.096721	.636984753	.411793086
WL	.258	.066564	.636984753	.564364554

SCALE CA	ESTIMATE	SQUARED CORRELATION	AVERAGE FOR CA	AVERAGE FOR SCALE
JS	.690	.476100	.493354098	.622632544
LA	.669	.447561	.493354098	.458698433
WH	.359	.128881	.493354098	.411793086
WL	.230	.052900	.493354098	.564364554

SCALE JS	ESTIMATE	SQUARED CORRELATION	AVERAGE FOR JS	AVERAGE FOR SCALE
LA	.694	.481636	.622632544	**.458698433**
WH	.462	.213444	.622632544	.411793086
WL	.329	.108241	.622632544	.564364554

SCALE LA	ESTIMATE	SQUARED CORRELATION	AVERAGE FOR LA	AVERAGE FOR SCALE
WH	.456	.298116	.458698433	.411793086
WL	.357	.127449	.458698433	.564364554

SCALE WH	ESTIMATE	SQUARED CORRELATION	AVERAGE FOR WH	AVERAGE FOR SCALE
WL	.717	.514089	**.411793086**	.564364554

Average variance extracted less than .50 is underlined; **average variance extracted proportions less than the squared correlation is in boldface.** GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface.

As indicated in Table 5.45, the average variance extracted for the General Sources of Stress sub-scale, the Role Ambiguity sub-scale, the Career Advancement sub-scale, the Lack of Autonomy sub-scale and the Work/Home Interface sub-scale are less than .50, suggesting that more variance in the indicators comprising these sub-scales was being explained by measurement error, as opposed to the constructs the scale items were designed to reflect. In addition, the average variance extracted for these five scales were less than the squared correlation. The finding that the average variance extracted for the RA sub-scale is less than the squared correlation for the RA and LA sub-scales means that the unique part of the RA latent variable is not sufficiently adequately measured. The RA items therefore do not successfully discriminate between the unique aspects of RA and LA. The same can be said for the REL and LA sub-scales as well as the LA and JS sub-scales and finally the WH and WL sub-scales. Therefore, the REL and LA sub-scales do not successfully discriminate between the unique aspects of REL and LA, the LA items do not successfully discriminate between the unique aspects of LA and JS and the WH items do not successfully discriminate between the unique aspects of WH and WL.

5.8.2 95% CONFIDENCE INTERVAL ESTIMATE: MALE SAMPLE

Calculation of the 95% confidence interval estimates involved examining the correlation coefficients between the latent variables of interest (**Φ**). Upon inspection of the results (Table 5.46), none of the 95% confidence interval estimates included unity, suggesting that there is sufficient evidence that all 36 null hypotheses, H_{0i}: $p_{pq} = 1$, hypothesising perfect correlations between the latent stress dimensions in the parameter, should be rejected. Therefore, there was sufficient evidence that the latent variables being measured by the SWSI were qualitatively distinct for the male sample. In addition to this, it should be noted that the correlations, for the male sample, were not excessively high (< .90).

Table 5.46

95% confidence interval estimate: male sample

	ESTIMATE	STANDARD ERROR ESTIMATE	LOWER LIMIT OF 95% CONFIDENCE INTERVAL	UPPER LIMIT OF 95% CONFIDENCE INTERVAL
GEN-RA	.594	.043	.503	.672
GEN-REL	.298	.053	.191	.398
GEN-TE	.250	.050	.150	.345
GEN-CA	.260	.054	.151	.362
GEN-JS	.424	.049	.323	.515
GEN-LA	.402	.050	.300	.495
GEN-WH	.600	.042	.511	.676
GEN-WL	.569	.049	.465	.657
RA-REL	.626	.041	.539	.700
RA-TE	.482	.047	.385	.569
RA-CA	.478	.050	.374	.570
RA-JS	.569	.047	.470	.654
RA-LA	.653	.046	.553	.734
RA-WH	.562	.049	.458	.650
RA-WL	.478	.051	.372	.572
REL-TE	.569	.044	.477	.649
REL-CA	.570	.040	.486	.643
REL-JS	.509	.045	.416	.592
REL-LA	.778	.028	.717	.827
REL-WH	.368	.051	.264	.463
REL-WL	.221	.052	.117	.320
TE-CA	.445	.049	.344	.536
TE-JS	.410	.054	.299	.510
TE-LA	.624	.046	.525	.706
TE-WH	.311	.053	.204	.411
TE-WL	.258	.052	.154	.357
CA-JS	.690	.036	.613	.754
CA-LA	.669	.041	.581	.742

CA-WH	.359	.051	.255	.455
CA-WL	.230	.055	.120	.335
JS-LA	.694	.040	.607	.764
JS-WH	.462	.045	.369	.546
JS-WL	.329	.054	.219	.430
LA-WH	.456	.047	.359	.543
LA-WL	.357	.052	.251	.454
WH-WL	.717	.032	.648	.774

GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface

5.8.3 SUMMARY OF DISCRIMINANT VALIDITY: MALE SAMPLE

The confidence interval estimate has a bearing on whether the latent variables that the items are measuring are qualitatively distinct. The average variance extracted versus the squared correlation has a bearing on whether the measures successfully capture the distinction that exists in the latent variables. Taking both tests into account, the results for the GEN, RA, CA, LA, and WH subscales were problematic as more variance in the indicators comprising the sub-scale is explained by measurement error, than by the constructs the sub-scale items were designed to reflect. Furthermore, the RA and LA sub-scales, REL and LA sub-scales, LA and JS sub-scales and WH and WL sub-scales do share variance however they do not correlate perfectly. They, therefore, each appear to have unique aspects. All the SWSI sub-scales can be described as qualitatively distinct, even though particular stress sub-scales do share more variance. Based on these results it was concluded that sufficient evidence were provided for discriminant validity in the male sample.

5.8.4 AVERAGE VARIANCE EXTRACTED VERSUS SHARED VARIANCE: FEMALE SAMPLE

The results for the average variance extracted versus shared variance, which compares the average variance extracted proportions for any two constructs with the square of the correlation estimate between these two constructs, are displayed in Table 5.47 for the female sample.

Table 5.47

Average variance extracted versus squared correlation: female sample

SCALE	ESTIMATE	SQUARED CORRELATION	AVERAGE FOR GEN	AVERAGE FOR SCALE
RA	.572	.327184	.525974771	.431797568
REL	.304	.092416	.525974771	.610514021
TE	.218	.047524	.525974771	.662454615
CA	.239	.057121	.525974771	.570686232

JS	.349	.121801	.525974771	.670631171
LA	.430	.184900	.525974771	.485438525
WH	.581	.337561	.525974771	.460797018
WL	.527	.277729	.525974771	.550624477

	ESTIMATE	**SQUARED CORRELATION**	**AVERAGE FOR RA**	**AVERAGE FOR SCALE**
REL	.561	.314721	.431797568	.610514021
TE	.471	.221841	.431797568	.662454615
CA	.435	.189225	.431797568	.570686232
JS	.477	.227529	.431797568	.670631171
LA	.597	.356409	.431797568	.485438525
WH	.416	.173056	.431797568	.460797018
WL	.402	.161604	.431797568	.550624477

	ESTIMATE	**SQUARED CORRELATION**	**AVERAGE FOR REL**	**AVERAGE FOR SCALE**
TE	.600	.360000	.610514021	.662454615
CA	.642	.412164	.610514021	.570686232
JS	.508	.258064	.610514021	.670631171
LA	.764	.583696	.610514021	**.485438525**
WH	.308	.094864	.610514021	.460797018
WL	.335	.112225	.610514021	.550624477

	ESTIMATE	**SQUARED CORRELATION**	**AVERAGE FOR TE**	**AVERAGE FOR SCALE**
CA	.567	.321489	.662454615	.570686232
JS	.419	.175561	.662454615	.670631171
LA	.641	.410881	.662454615	.485438525
WH	.334	.111556	.662454615	.460797018

WL	.370	.136900	.662454615	.550624477

	ESTIMATE	**SQUARED CORRELATION**	**AVERAGE FOR CA**	**AVERAGE FOR SCALE**
JS	.724	.524176	.570686232	.670631171
LA	.729	.531441	.570686232	**.485438525**
WH	.326	.106276	.570686232	.460797018
WL	.286	.081796	.570686232	.550624477

	ESTIMATE	**SQUARED CORRELATION**	**AVERAGE FOR JS**	**AVERAGE FOR SCALE**
LA	.699	.488601	.670631171	**.485438525**
WH	.409	.167281	.670631171	.460797018
WL	.322	.103684	.670631171	.550624477

	ESTIMATE	**SQUARED CORRELATION**	**AVERAGE FOR LA**	**AVERAGE FOR SCALE**
WH	.497	.247009	.485438525	.460797018
WL	.442	.195364	.485438525	.550624477

	ESTIMATE	**SQUARED CORRELATION**	**AVERAGE FOR WH**	**AVERAGE FOR SCALE**
WL	.648	.419904	.460797018	.550624477

Average variance extracted less than .50 is underlined; **average variance extracted proportions less than the squared correlation is in bold**. GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface.

The results, as indicated in Table 5.47, show that the average variance extracted for the Role Ambiguity (RA) sub-scale, the Lack of Autonomy (LA) sub-scale and the Work/Home Interface sub-scales is less than .50, which suggests that more variance in the indicators comprising these sub-scales is being explained by measurement error, as opposed to the constructs that the sub-scale items were designed to reflect. In addition, the average variance extracted for three sub-scales is less than the squared correlation. The finding that the average variance extracted for the LA sub-scale is less than the squared correlation for the LA and REL sub-scales suggests that the unique part

of the LA latent variable is not sufficiently adequately measured. The LA items therefore do not successfully discriminate between the unique aspects of LA and REL. The same can be said for the LA and CA sub-scales as well as the LA and JS sub-scales. Therefore, the LA measures do not successfully discriminate between the unique aspect of LA and CA as well as LA and JS.

5.8.5 95% CONFIDENCE INTERVAL ESTIMATE: FEMALE SAMPLE

Calculation of the 95% confidence interval estimates involved examining the correlation coefficients between the latent variables of interest ($\boldsymbol{\Phi}$). Upon inspection of the results (Table 5.48), none of the 95% confidence interval estimates included unity, thereby suggesting that there was sufficient evidence that all 36 null hypotheses H_{0i}: $p_{pq} = 1$, hypothesising perfect correlations between the 9 latent stress dimensions in the parameter, should be rejected. Therefore, there is sufficient evidence that the latent stress dimensions being measured by the SWSI items are qualitatively distinct for the female sample. In addition to this, it should be noted that the correlations, in the female sample, were not excessively high (< .90).

Table 5.48

95% confidence interval estimate: female sample

	ESTIMATE	STANDARD ERROR ESTIMATE	LOWER LIMIT OF 95% CONFIDENCE INTERVAL	UPPER LIMIT OF 95% CONFIDENCE INTERVAL
GEN-RA	.572	.041	.486	.647
GEN-REL	.304	.050	.203	.399
GEN-TE	.218	.048	.122	.310
GEN-CA	.239	.049	.141	.332
GEN-JS	.349	.048	.252	.439
GEN-LA	.430	.048	.331	.519
GEN-WH	.581	.042	.493	.657
GEN-WL	.527	.052	.418	.621
RA-REL	.561	.041	.475	.636
RA-TE	.471	.045	.378	.554
RA-CA	.435	.045	.343	.519
RA-JS	.477	.044	.386	.559
RA-LA	.597	.042	.508	.673
RA-WH	.416	.047	.320	.504
RA-WL	.402	.050	.300	.495
REL-TE	.600	.041	.514	.674
REL-CA	.642	.035	.568	.706
REL-JS	.508	.043	.419	.587
REL-LA	.764	.026	.708	.810
REL-WH	.308	.055	.197	.412
REL-WL	.335	.049	.236	.427
TE-CA	.567	.042	.479	.644
TE-JS	.419	.050	.316	.512

TE-LA	.641	.037	.563	.708
TE-WH	.334	.054	.224	.435
TE-WL	.370	.049	.270	.462
CA-JS	.724	.032	.655	.781
CA-LA	.729	.032	.660	.786
CA-WH	.326	.058	.208	.435
CA-WL	.286	.055	.175	.390
JS-LA	.699	.035	.624	.761
JS-WH	.409	.053	.300	.507
JS-WL	.322	.052	.217	.420
LA-WH	.497	.050	.393	.589
LA-WL	.442	.051	.337	.536
WH-WL	.648	.037	.570	.715

GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface

5.8.6 SUMMARY OF DISCRIMINANT VALIDITY: FEMALE SAMPLE

Taking the results of both tests into account, the results for the RA, LA, and WH sub-scales are problematic, as more variance in the indicators comprising the sub-scale is explained by measurement error than by the constructs the sub-scale items were designed to reflect. Furthermore, the results revealed that the LA and REL sub-scales, LA and CA sub-scales, and LA and JS sub-scales do share variance. However, these sub-scales do not correlate perfectly, and therefore each sub-scale seems to measure unique elements, not measured by the other sub-scale. All the SWSI sub-scales can be described as qualitatively distinct, even though particular stress sub-scales do share more variance. Sufficient evidence was provided for discriminant validity in the female sample.

5.9 CONFIGURAL INVARIANCE

The finding of satisfactory model fit for both gender samples independently, justified further measurement invariance and equivalence analyses. The next step involved investigating configural invariance.

Operational hypothesis 2 was tested by testing H02: RMSEA ≤ .05.

The multi-group SWSI measurement model in which the structure was constrained to be equal across gender groups, but with all the model parameters freely estimated, was fitted to the male (N=460) and female samples (N=460) simultaneously. A visual representation of the fitted multi-group configural invariance measurement model is shown in Figure 5.7.

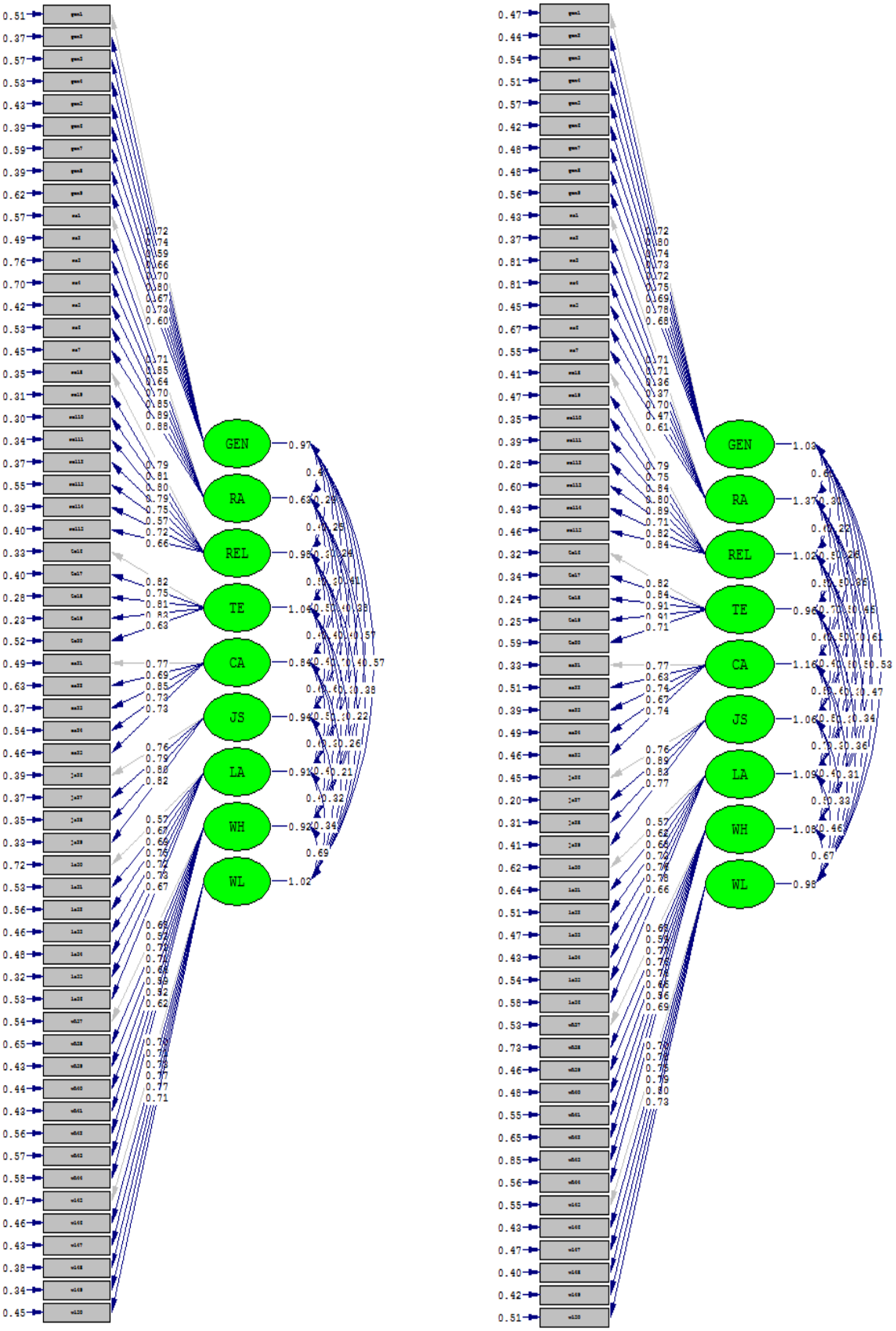

Figure 5.7. Representation of the fitted multi-group SWSI configural invariance measurement model for the male and female samples respectively

5.9.1 MEASUREMENT MODEL FIT INDICES

The multi-group configural invariance measurement model converged in 42 iterations. The spectrum of fit statistics calculated by LISREL 9.00 is shown in Table 5.49.

Table 5.49

Goodness of fit statistics for the multi-group SWSI configural invariance measurement model

Degrees of Freedom = 3232
Minimum Fit Function Chi-Square = 8244.029 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 8900.462 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 7412.868 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 4180.868
90 Percent Confidence Interval for NCP = (3934.059 ; 4435.197)

Minimum Fit Function Value = 8.980
Population Discrepancy Function Value (F0) = 4.554
90 Percent Confidence Interval for F0 = (4.285 ; 4.831)
Root Mean Square Error of Approximation (RMSEA) = 0.0531
90 Percent Confidence Interval for RMSEA = (0.0515 ; 0.0547)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000

Expected Cross-Validation Index (ECVI) = 9.003
90 Percent Confidence Interval for ECVI = (8.606 ; 9.152)
ECVI for Saturated Model = 3.856
ECVI for Independence Model = 161.365

Chi-Square for Independence Model with 3422 Degrees of Freedom = 148015.476
Independence AIC = 148251.476
Model AIC = 8264.868
Saturated AIC = 7080.000
Independence CAIC = 148938.752
Model CAIC = 10746.051
Saturated CAIC = 27698.283

Normed Fit Index (NFI) = 0.950
Non-Normed Fit Index (NNFI) = 0.969
Parsimony Normed Fit Index (PNFI) = 0.897
Comparative Fit Index (CFI) = 0.971
Incremental Fit Index (IFI) = 0.971
Relative Fit Index (RFI) = 0.947

Critical N (CN) = 424.773

Upon first inspection of Table 5.49, the degrees of freedom were indicated to be 3232. This corresponds with the calculations in section 4.5.1.2 and provides evidence that the model was specified correctly.

The exact fit null hypothesis was not tested, therefore the following close fit null hypothesis, which explicitly assumes that the measurement model only approximates the processes that operated in reality to create the observed covariance matrix (Browne & Cudeck, 1993), was tested:

$$H_{02}: \text{RMSEA} \leq .05$$

$$H_{a2}: \text{RMSEA} > .05$$

Similar to the cut-off values that are applicable the single group model, for the multi-group models RMSEA values below .05 are generally regarded as indicative of good model fit, values above .05 but less than .08 as indicative of reasonable fit, values greater than or equal to .08 but less than .10 indicative of mediocre fit and values exceeding .10 are generally regarded as indicative of poor fit. The results for this analysis (Table 5.49) revealed a RMSEA value of .0531, thus indicating that the measurement model showed reasonable to good fit in the sample. The p-value for Test of Close Fit (H_{02}: RMSEA < .05) was 1.00, therefore the close fit null hypothesis H_{02}: RMSEA ≤ .05 was not rejected (p > .05). The position that the measurement model shows close fit in the parameter was therefore a tenable position. The multi-group measurement model is therefore a plausible explanation for the observed covariance matrices in that the configural invariance multi-group model approximately reproduces the observed covariance matrices, but not perfectly.

The confidence interval, which assists in assessing the precision of the fit statistics, for the RMSEA index was (.0515 ; .0547). The fact that the interval was small indicates a higher level of precision in reflecting the model fit in the population (Byrne, 2001).

5.9.2 DECISION ON CONFIGURAL INVARIANCE

Upon fitting the multi-group configural invariance measurement model, it was established that the multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, but with no freed parameters constrained to be equal across groups, displayed close fit when fitted to the samples simultaneously in a multi-group analysis. This finding may be interpreted to signify a lack of construct bias. It may be concluded that the different gender groups used the same conceptual frame of reference when they responded to the items. Based on the results of the analysis it is clear that the SWSI occupational stress instrument reflects the same underlying constructs across gender groups.

Configural invariance is a prerequisite for evaluating further aspects of measurement invariance and measurement equivalence. The SWSI displayed configural invariance, therefore other tests of measurement invariance and equivalence were allowed.

5.10 WEAK INVARIANCE

Upon finding (a) acceptable single-group measurement model fit on both gender samples independently and (b) support for multi-group configural invariance, weak invariance was tested.

Operational hypothesis 3 was tested by testing H_{03}: RMSEA ≤ .05.

The multi-group SWSI measurement model, in which the structure of the model was constrained to be equal and the slopes of the regression of the indicator variables on the latent variables were constrained to be equal, but all other parameters were estimated freely across the gender groups, was fitted to the male (N=460) and female (N=460) samples. A visual representation of the fitted multi-group weak invariance model is shown in Figure 5.8.

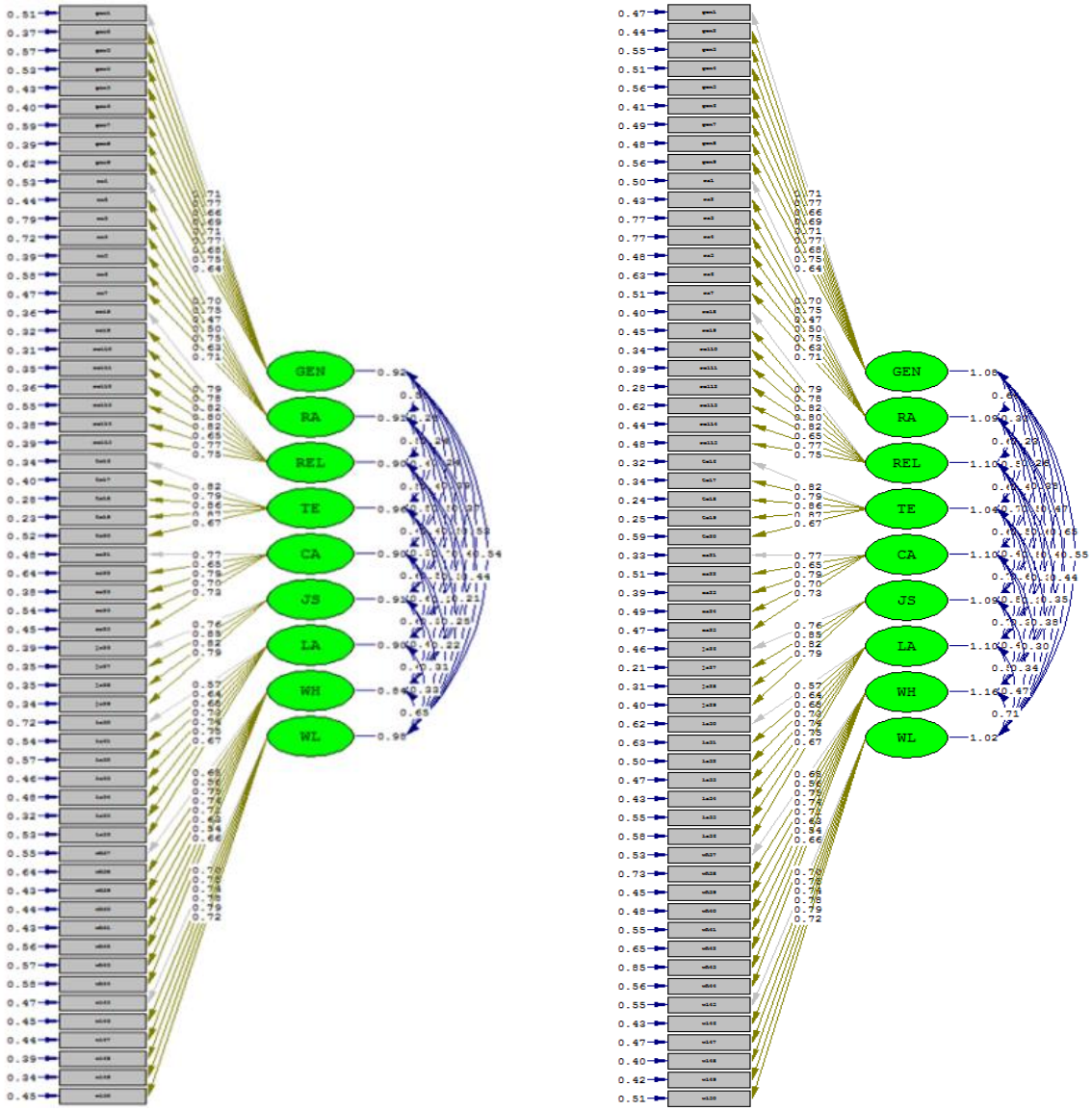

Figure 5.8. Representation of the fitted multi-group SWSI weak invariance measurement model for the male and female sample respectively

5.10.1 MEASUREMENT MODEL FIT INDICES

The weak invariance measurement model converged in 36 iterations. The spectrum of fit statistics calculated by LISREL 9.00 is shown in Table 5.50.

Table 5.50

Goodness of fit statistics for the multi-group SWSI weak invariance measurement model

Degrees of Freedom = 3282
Minimum Fit Function Chi-Square = 8334.509 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 9021.990 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 7540.733 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 4258.733
90 Percent Confidence Interval for NCP = (4009.707 ; 4515.274)

Minimum Fit Function Value = 9.079
Population Discrepancy Function Value (F0) = 4.639
90 Percent Confidence Interval for F0 = (4.368 ; 4.919)
Root Mean Square Error of Approximation (RMSEA) = 0.0532
90 Percent Confidence Interval for RMSEA = (0.0516 ; 0.0547)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000

Expected Cross-Validation Index (ECVI) = 9.033
90 Percent Confidence Interval for ECVI = (8.634 ; 9.184)
ECVI for Saturated Model = 3.856
ECVI for Independence Model = 161.365

Chi-Square for Independence Model with 3422 Degrees of Freedom = 148015.476
Independence AIC = 148251.476
Model AIC = 8292.733
Saturated AIC = 7080.000
Independence CAIC = 148938.752
Model CAIC = 10482.698
Saturated CAIC = 27698.283

Normed Fit Index (NFI) = 0.949
Non-Normed Fit Index (NNFI) = 0.969
Parsimony Normed Fit Index (PNFI) = 0.910
Comparative Fit Index (CFI) = 0.971
Incremental Fit Index (IFI) = 0.971
Relative Fit Index (RFI) = 0.947

Critical N (CN) = 423.849

The degrees of freedom for this model were 3282. This corresponds with the calculations in Table 4.1 and provides evidence that the model was specified correctly.

The exact fit null hypothesis was not tested, therefore, the following close fit null hypothesis, which explicitly assumes that the measurement model only approximates the processes that operated in reality to create the observed covariance matrix (Browne & Cudeck, 1993), was tested:

$$H_{03}: \text{RMSEA} \leq .05$$

$$H_{a3}: \text{RMSEA} > .05$$

A RMSEA value of .0532 was obtained, indicating that the measurement model showed reasonable to good fit in the sample. The p-value for Test of Close Fit (H_{03}: RMSEA < .05) was 1.00, therefore the close fit null hypothesis H_{03}: RMSEA ≤ .05 was not rejected. The position that the multi-group weak invariance measurement model showed close fit was therefore permissible. The multi-group measurement model may therefore be considered a plausible explanation for the observed covariance matrices in that the multi-group weak invariance model could approximately reproduce the observed covariance matrices, but not perfectly.

The RMSEA confidence interval, which assists in assessing the precision of the fit statistics, for this index was (.0516 ; .0547). The small interval further reflected a higher level of precision of model fit in the population (Byrne, 2001).

5.10.1 DECISION ON WEAK INVARIANCE

Upon fitting the multi-group weak invariance measurement model, it was established that the multi-group measurement model in which the structure of the model were constrained to be the same across gender groups, and in which all parameters were estimated freely across the gender samples, but for the slopes of the regression of the indicator variables on the latent variables that were constrained to be equal, demonstrated close fit when fitted to the gender samples simultaneously in a multi-group analysis. Weak invariance was thus displayed, which implies that the position that the slopes of the regression of the items on the latent variables they represent are the same across the samples, was a tenable one. Item content was therefore being perceived and interpreted the same across gender samples and support was provided for the position that the SWSI items operate in approximately the same way across gender samples in the way they reflect the underlying latent stress variables they were meant to reflect. Therefore lack of non-uniform item bias has been established as a tenable position. The position that the slopes of the regression of the items on the latent variables they represent are not the same across the samples may be a more convincing position. To investigate this latter possibility metric equivalence was tested.

5.11 METRIC EQUIVALENCE

A finding of weak invariance allowed for the testing of metric equivalence. The test of metric equivalence via multi-group confirmatory factor analysis using LISREL determines whether the multi-group (weak invariance) measurement model in which the Λ^x is constrained to be equal across groups fits (a) statistically significantly better or (b) practically significantly better than a multi-group (configural invariance) measurement model in which all model parameters are estimated freely. The decision on whether the multi-group SWSI weak invariance measurement model shows metric equivalence is based on the question whether the multi-group weak invariance model fits practically significantly poorer than the multi-group configural invariance model. The results of the test whether the multi-group weak invariance model fits statistically significantly poorer than the multi-group configural invariance model are nonetheless also provided.

The results of the test of the statistical significance of the difference in multi-group measurement model fit are shown in Table 5.51.

Table 5.51

Statistical significance of the scaled chi-squared difference statistic: a test of metric equivalence

HYPOTHESIS	SATORRA-BENTLER CHI SQUARE	NORMAL THEORY CHI-SQUARE	df	Cd	SCALED DIFFERENCE IN S-B CHI-SQUARE	PROB S-B CHI-SQUARE DIFF	PROB SCALED S-B CHI-SQUARE DIFF
H_a: CONFIGURAL INVARIANCE MODEL	7412.868	8900.462	3232				
H_{03}: WEAK INVARIANCE MODEL	7540.733	9021.99	3282				
DIFF (H_{03}-H_a) METRIC EQUIVALENCE	**127.865**		**50**	**0.922155705**	**131.7868548**	**9.49386E-09**	**2.71356E-09**

If the probability of observing the scaled chi-squared difference in a multi-group sample under the null hypothesis of no difference in fit between the weak and configural invariance models in the parameter (H_0:$\chi^2 = 0$) is smaller than or equal to .05, the null hypothesis is rejected in favour of the hypothesis that the fit of the multi-group configural invariance and multi-group weak invariance models differ in the parameter. As can be seen in Table 5.51, the null hypothesis of no difference in

model fit in the parameter was rejected ($p < .05$) and this implies lack of equivalence of factor loadings across the two gender groups (i.e. lack of metric equivalence) when using statistical significance as the yardstick of equivalence. Cheung and Rensvold (2002), however, argue against the sole reliance on statistical significance to decide on measurement equivalence and also recommend the use of alternative indices to reflect on the practical significance in multi-group model fit.

The results for the test of practical significance of the difference in multi-group measurement model fit are shown in Table 5.52.

Table 5.52

Practical significance of the CFI, Gamma Hat and MacDonald difference statistics: a test of metric equivalence

MODEL	CFI	Γ_1	Mc
CONFIGURAL INVARIANCE H_a MODEL	0.971	0.9283444	0.10259152
WEAK INVARIANCE H_{03} MODEL	0.971	0.927104449	0.098322735
DIFF [H_{03}-H_a; TEST OF METRIC EQUIVALENCE]	**0**	**-0.001239952**	**-0.004268785**

CFI= Comparative Fit Index; Γ_1= Gamma Hat fit index; Mc= McDonald Non-centrality index

Metric equivalence may be concluded if the weak invariance multi-group measurement model does not fit practically significantly poorer than the configural invariance multi-group measurement model. The difference in fit can be considered practically insignificant if a change of -0.01 or less in the CFI fit index, a change of -0.001 or less in the Gamma Hat fit index ($\boldsymbol{\Gamma 1}$) and a change of -0.02 or less in the Mcdonald Non-centrality index (Cheung & Rensvold, 2002) is observed between the weak invariance multi-group measurement model and configural invariance multi-group measurement model. As indicated in Table 5.52, the change in CFI and Mc was less than the critical thresholds, however Gamma Hat fit index was marginally greater than the critical threshold of -0.001. In terms of the decision-rule specified in Chapter 4, metric equivalence could therefore not be concluded. This suggests that a multi-group measurement model in which the structure of the model is constrained to be the same across gender groups and in which all parameters are estimated freely across the gender samples, but for the slopes of the regression of the indicator variables on the latent variables, fits practically poorer than a multi-group measurement model in which the

structure of the model is constrained to be the same across gender groups but all parameters are estimated freely. This in turn suggests that the slope parameter estimates for one or more items differ practically significantly across the two gender groups.

5.11.1 DECISION ON THE RESULTS OF WEAK INVARIANCE AND METRIC EQUIVALENCE

Upon fitting the weak invariance multi-group measurement model, the results indicated acceptable model fit, permitting the conclusion that lack of non-uniform item bias was a tenable position to hold with regards to the SWSI. However, when comparing the weak invariance multi-group measurement model to the configural invariance multi-group measurement model (i.e. one with fewer constraints), metric equivalence was not supported as the difference on the Gamma Hat fit index was greater than the critical threshold of -0.001. Therefore, although the weak invariance multi-group measurement model, in which the structure of the model is constrained to be the same across gender groups and in which all parameters are estimated freely across the gender samples, but for the slopes of the regression of the indicator variables on the latent variables, fitted the data closely, the results suggested that the configural invariance multi-group measurement model fitted the data practically significantly better. The position of the presence of non-uniform bias was therefore a more tenable position than the position of the presence of no non-uniform bias (although the latter position was also permissible).

The inadequate support for metric equivalence indicates that one or more factor loadings differ across groups. The possibility of partial metric equivalence was therefore explored. This required additional tests to determine the sources of non-invariance. Lack of metric equivalence could be due to non-invariant construct(s) or non-invariant item(s) used to reflect invariant construct(s) (Milfont, Duckitt & Cameron, 2006). Configural invariance has already been established. The problem therefore had to lie with the differences in the responsiveness of one or more items to changes in the underlying latent stress dimension they reflect across gender samples. Accurate identification of non-invariant constructs and items is a prerequisite for meaningful cross-group comparison (Cheung & Rensvold, 2002).

5.12 PARTIAL METRIC EQUIVALENCE

A lack of support for metric equivalence resulted in the need for additional tests to determine the source of non-equivalence, therefore testing for partial metric equivalence was conducted next. All non-invariant items must be correctly identified (Cheung & Rensvold, 1999).

The procedure followed involved (a) testing for partial metric equivalence per sub-scale, (b) examining the factor loadings of the configural invariance model for those sub-scales where a lack of partial metric equivalence was found, to identify the greatest difference between gender groups (i.e. the most dissimilar factor loadings), and (c) testing for partial metric equivalence when lifting the factor loading equality constraint one item at a time, starting with the item with the most dissimilar loadings across the two gender samples.

The SWSI measurement model consists of multiple latent stress dimensions, each measured by a sub-scale of items. The item/items for which the regression of X_i on ξ_j differ(s) across gender samples can be limited to a specific sub-scale or can be scattered across two or more sub-scales. To narrow the search for the non-invariant items down to specific sub-scales each of the stress sub-scales in the model were examined for invariance (Cheung & Rensvold, 1999). A separate multi-group model was estimated in which the factor loadings associated with a specific construct (i.e. sub-scale) were constrained to be equal across gender groups, while the loadings associated with the other stress constructs (i.e. sub-scales) were freely estimated. The sub-scale in which statistical and/or practical significance was displayed indicated that at least one of the items within that sub-scale could be considered as non-invariant (Cheung & Rensvold, 1999). All non-invariant sub-scales were noted, and their items examined for invariance.

Upon fitting each multi-group partial weak invariance model per stress sub-scale, the degrees of freedom were examined. The degrees of freedom tallied with the calculations[26], therefore it could be assumed that the models were estimated correctly. The statistical significance of the scaled Satorra-Bentler difference statistic for each partial weak invariance multi-group model is shown in Table 5.53.

[26] The degrees of freedom calculations for each partial weak invariance model were calculated separately by the researcher in order to ensure each model was specified correctly.

Table 5.53

Statistical significance of the scaled chi-squared difference statistic: a test of partial metric equivalence (constructs)

HYPOTHESIS	SATORRA-BENTLER CHI SQUARE	NORMAL THEORY CHI-SQUARE	df	Cd	SCALED DIFFERENCE IN S-B CHI-SQUARE	PROB S-B CHI-SQUARE DIFF	PROB SCALED S-B CHI-SQUARE DIFF
H_a CONFIGURAL INVARIANCE MODEL	7412.868	8900.462	3232				
PARTIAL WEAK INVARIANCE GEN SCALE H_{03a} MODEL	7430.979	8916.85	3240				
PARTIAL WEAK INVARIANCE RA SCALE H_{03b} MODEL	7453.731	8947.624	3238				
PARTIAL WEAK INVARIANCE REL SCALE H_{03c} MODEL	7445.803	8937.317	3239				
PARTIAL WEAK INVARIANCE TE SCALE H_{03d} MODEL	7416.646	8902.151	3236				
PARTIAL WEAK INVARIANCE CA SCALE H_{03e} MODEL	7420.406	8906.795	3236				
PARTIAL WEAK INVARIANCE JS SCALE H_{03f} MODEL	7424.031	8913.884	3235				
PARTIAL WEAK INVARIANCE LA SCALE H_{03g} MODEL	7419.626	8903.437	3238				
PARTIAL WEAK INVARIANCE WH SCALE H_{03h} MODEL	7421.293	8904.265	3239				
PARTIAL WEAK INVARIANCE WL SCALE H_{03i} MODEL	7418.649	8902.59	3237				
DIFF (H_{03a}-H_a) PARTIAL METRIC EQUIVALENCE	**18.111**		**8**	**0.908686976**	**18.03481335**	**0.020409526**	**0.020966995**
DIFF (H_{03b}-H_a) PARTIAL METRIC EQUIVALENCE	**40.863**		**6**	**1.06302075**	**44.36602015**	**3.08129E-07**	**6.254E-08**
DIFF (H_{03c}-H_a) PARTIAL METRIC EQUIVALENCE	**32.935**		**7**	**1.033552229**	**35.65857533**	**2.72249E-05**	**8.4077E-06**
DIFF (H_{03d}-H_a) PARTIAL METRIC EQUIVALENCE	**3.778**		**4**	**0.890112161**	**1.897513678**	**0.43688312**	**0.754601723**
DIFF (H_{03e}-H_a) PARTIAL METRIC EQUIVALENCE	**7.538**		**4**	**0.904383027**	**7.002563968**	**0.11004528**	**0.135752793**
DIFF (H_{03f}-H_a) PARTIAL METRIC EQUIVALENCE	**11.163**		**3**	**1.203413718**	**11.15327156**	**0.010876348**	**0.010925303**
DIFF (H_{03g}-H_a) PARTIAL METRIC EQUIVALENCE	**6.758**		**6**	**0.826879106**	**3.59786573**	**0.343808248**	**0.730906831**
DIFF (H_{03h}-H_a) PARTIAL METRIC EQUIVALENCE	**8.425**		**7**	**0.807082879**	**4.712031565**	**0.296612486**	**0.695057378**
DIFF (H_{03i}-H_a) PARTIAL METRIC EQUIVALENCE	**5.781**		**5**	**0.780653233**	**2.725922227**	**0.328115184**	**0.742150983**

GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface.

If the probability of observing the scaled chi-squared difference in a multi-group sample under the null hypothesis of no difference in fit between the partial weak and configural invariance models in the parameter (H_0:$\chi^2 = 0$) is smaller than or equal to .05, the null hypothesis is rejected in favour of the hypothesis that the fit of the multi-group configural invariance and multi-group partial weak invariance models differ in the parameters. As can be seen in Table 5.53, the null hypothesis of no difference in model fit in the parameter was rejected ($p < .05$) for the GEN, RA, REL and JS sub-scales. This implies lack of equivalence of factor loadings across the two gender groups for these sub-scales (i.e. lack of partial metric equivalence), when using statistical significance as the yardstick of equivalence. As is evident from Table 5.53, the null hypothesis of no difference in model fit in the parameters was not rejected ($p > .05$) for the TE, CA, LA, WH and WL subscales.

The results for the test of practical significance of the difference in multi-group measurement model fit are shown in Table 5.54.

Table 5.54

Practical significance of the CFI, Gamma Hat and MacDonald difference statistic: a test of partial metric equivalence per scale

MODEL	CFI	Γ_1	Mc
CONFIGURAL INVARIANCE H_a MODEL	0.971	0.9283444	0.10259152
PARTIAL WEAK INVARIANCE GEN SCALE H_{03a} MODEL	0.971	0.928505107	0.103157328
PARTIAL WEAK INVARIANCE RA SCALE H_{03b} MODEL	0.971	0.927789659	0.100660682
PARTIAL WEAK INVARIANCE REL SCALE H_{03c} MODEL	0.971	0.927920985	0.101114676
PARTIAL WEAK INVARIANCE TE SCALE H_{03d} MODEL	0.971	0.9283444	0.10259152
PARTIAL WEAK INVARIANCE CA SCALE H_{03e} MODEL	0.971	0.928285975	0.102386542
PARTIAL WEAK INVARIANCE JS SCALE H_{03f} MODEL	0.971	0.928212954	0.102130895
PARTIAL WEAK INVARIANCE LA SCALE H_{03g} MODEL	0.971	0.928329793	0.102540237
PARTIAL WEAK INVARIANCE WH SCALE H_{03h} MODEL	0.971	0.928315187	0.10248898
PARTIAL WEAK INVARIANCE WL SCALE H_{03i} MODEL	0.971	0.928329793	0.102540237
DIFF [H_{03a}-H_a; TEST OF PARTIAL GEN METRIC EQUIVALENCE]	**0**	**0.000160707**	**0.000565808**
DIFF [H_{03b}-H_a; TEST OF RA PARTIAL METRIC EQUIVALENCE]	**0**	**-0.000554741**	**-0.001930838**
DIFF [H_{03c}-H_a; TEST OF REL PARTIAL METRIC EQUIVALENCE]	**0**	**-0.000423415**	**-0.001476844**
DIFF [H_{03d}-H_a; TEST OF TE PARTIAL METRIC EQUIVALENCE]	**0**	**0**	**0**
DIFF [H_{03e}-H_a; TEST OF CA PARTIAL METRIC EQUIVALENCE]	**0**	**-5.8425E-05**	**-0.000204978**
DIFF [H_{03f}-H_a; TEST OF JS PARTIAL METRIC EQUIVALENCE]	**0**	**-0.000131446**	**-0.000460625**
DIFF [H_{03g}-H_a; TEST OF LA PARTIAL METRIC EQUIVALENCE]	**0**	**-1.46069E-05**	**-5.12829E-05**
DIFF [H_{03h}-H_a; TEST OF WH PARTIAL METRIC EQUIVALENCE]	**0**	**-2.92134E-05**	**-0.00010254**
DIFF [H_{03i}-H_a; TEST OF WL PARTIAL METRIC EQUIVALENCE]	**0**	**-1.46069E-05**	**-5.12829E-05**

CFI= Comparative Fit Index; Γ_1= Gamma Hat fit index; Mc= McDonald Non-centrality index

Partial metric equivalence for each sub-scale may be concluded if the partial weak invariance multi-group measurement model for each sub-scale does not fit practically significantly poorer than the configural invariance multi-group measurement model. The difference in fit can be considered practically insignificant if a change of -0.01 or less in the CFI fit index, a change of -0.001 or less in the Gamma Hat fit index (**Γ1**) and a change of -0.02 or less in the Mcdonald Non-centrality index (Cheung & Rensvold, 2002) is observed between the partial weak invariance multi-group measurement model for each sub-scale and configural invariance multi-group measurement model. As indicated in Table 5.54, the change in CFI, Γ1 and Mc was less than the critical thresholds for all the sub-scales. Partial metric equivalence could, therefore, be concluded for each sub-scale.

Adequate support was not provided for metric equivalence when the SWSI weak invariance multi-group measurement model, as a whole, was compared to the SWSI configural invariance multi-group measurement model. However, support for partial metric equivalence for each sub-scale was found. This anomaly is most likely due to the result of the fact that the lack of metric equivalence on all the items collectively is probably the result of the cumulative effect of small differences occurring across scales where those occurring in a single sub-scale are not sufficient to promote lack of partial metric equivalence.

Due to none of the sub-scales displaying lack of partial metric equivalence when using practical significance as the yardstick of equivalence, non-invariant sub-scales could not be noted and their items examined for invariance. Therefore, all items had to be considered for possible non-invariance. The differences between the male and female factor loadings in the Lambda X common metric completely standardised solution from the configural invariance model was calculated and the absolute values rank ordered from high to low in Table 5.55 in order to identify the items with the most dissimilar factor loadings.

Table 5.55

Lambda_X_difference

	Configural invariance: Lambda X		Lambda_X_diff	Rank order from largest to smallest	
	Male	**Female**	**Difference**	**Item**	**Difference**
GEN1	0.715	0.702	0.013	**LA35**	**0.433**
GEN2	0.738	0.816	-0.078	**JS27**	**0.392**
GEN3	0.586	0.746	-0.221	**RA6**	**0.341**
GEN4	0.659	0.807	-0.148	TE18	0.327
GEN5	0.701	0.675	0.026	RA4	0.311
GEN6	0.799	0.685	0.114	REL12	0.311
GEN7	0.672	0.759	-0.087	TE20	0.296
GEN8	0.727	0.734	-0.007	TE17	0.289
GEN9	0.595	0.693	-0.098	TE19	0.286
RA1	0.708	0.766	-0.058	CA21	0.28
RA2	0.854	0.782	0.072	REL10	0.269
RA3	0.64	0.43	0.21	WH44	0.265
RA4	0.701	0.39	0.311	REL15	0.259
RA5	0.848	0.695	0.153	CA25	0.25
RA6	0.893	0.552	0.341	REL8	0.248
RA7	0.882	0.675	0.207	WL49	0.245
REL8	0.786	1.034	-0.248	REL14	0.243
REL9	0.811	0.935	-0.124	LA34	0.238
REL10	0.804	1.073	-0.269	WH39	0.232
REL11	0.79	0.953	-0.163	WL46	0.232
REL12	0.746	1.057	-0.311	JS26	0.222
REL13	0.573	0.745	-0.172	GEN3	0.221
REL14	0.718	0.961	-0.243	JS28	0.216
REL15	0.664	0.923	-0.259	CA22	0.211
TE16	0.82	0.996	-0.176	RA3	0.21
TE17	0.754	1.043	-0.289	WL47	0.208
TE18	0.814	1.141	-0.327	RA7	0.207
TE19	0.832	1.118	-0.286	WH38	0.207
TE20	0.629	0.925	-0.296	WH40	0.191
CA21	0.77	1.05	-0.28	WH37	0.186
CA22	0.691	0.902	-0.211	TE16	0.176
CA23	0.853	1.009	-0.156	WL45	0.174
CA24	0.728	0.849	-0.121	REL13	0.172
CA25	0.732	0.982	-0.25	REL11	0.163
JS26	0.76	0.982	-0.222	CA23	0.156
JS27	0.791	1.183	-0.392	RA5	0.153
JS28	0.8	1.016	-0.216	LA36	0.153
JS29	0.818	0.908	-0.09	GEN4	0.148
LA30	0.572	0.637	-0.065	WL50	0.146
LA31	0.671	0.711	-0.04	WL48	0.131
LA32	0.69	0.804	-0.114	WH41	0.126
LA33	0.749	0.869	-0.12	REL9	0.124

LA34	0.717	0.955	-0.238	CA24	0.121
LA35	0.727	1.16	-0.433	LA33	0.12
LA36	0.674	0.827	-0.153	GEN6	0.114
WH37	0.682	0.868	-0.186	LA32	0.114
WH38	0.525	0.732	-0.207	WH42	0.105
WH39	0.719	0.951	-0.232	GEN9	0.098
WH40	0.706	0.897	-0.191	WH43	0.093
WH41	0.663	0.789	-0.126	JS29	0.09
WH42	0.585	0.69	-0.105	GEN7	0.087
WH43	0.523	0.43	0.093	GEN2	0.078
WH44	0.617	0.882	-0.265	RA2	0.072
WL45	0.701	0.875	-0.174	LA30	0.065
WL46	0.709	0.941	-0.232	RA1	0.058
WL47	0.734	0.942	-0.208	LA31	0.04
WL48	0.774	0.905	-0.131	GEN5	0.026
WL49	0.771	1.016	-0.245	GEN1	0.013
WL50	0.707	0.853	-0.146	GEN8	0.007

GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface. **Items with non-invariant factor loadings are in boldface.**

The process to identify non-invariant items continued with the performance of a series of tests. Using the rank ordered absolute values in Table 5.55, a multi-group model was estimated in which the item with the most dissimilar factor loading between gender groups was unconstrained from equality. Therefore, under the partial weak invariance multi-group measurement model that is fitted in order to test for partial metric equivalence, all factor loadings of items were constrained to be equal across groups, but for the item that had the most dissimilar factor loading in the fully unconstrained solution. The non-invariant loadings are permitted to be freely estimated (Cheung & Rensvold, 1999). In this case, the first partial weak invariance multi-group measurement model that was estimated was one in which the structure of the model was constrained to be the same across gender groups and in which all parameters were estimated freely across the gender samples, but for 58 of the slopes of the regression of the indicator variables on the latent variables that were constrained to be equal. The slope of the regression of the indicator variable LA35 was not constrained to be equal, and therefore allowed to vary, as this item displayed the largest factor loading difference between the gender groups. If the difference in fit between the partial weak invariance and configural invariance models were found to be statistically and/or practically insignificant, then partial equivalence would be obtained, and the item whose slope parameter was estimated freely across samples, would be considered as non-invariant. Under a finding of a lack of partial metric equivalence the process would be repeated by now also freeing the item that had the second most dissimilar factor loading between gender groups in the fully unconstrained solution in the partial weak invariance model. This process was continued until the difference in fit between the

partial weak invariance and configural invariance models was found to be practically insignificant. This was interpreted to signify that all non-invariant items had been identified. Given this procedure, the next partial weak invariance multi-group measurement model that was estimated allowed for the slopes of the regression of the indicators variables LA35 and JS27 to vary. Identification of all non-invariant items would result in the attainment of practical insignificance, thereby indicating that a partial weak invariance multi-group measurement model in which the non-invariant items are allowed to vary, does not fit practically significantly poorer than the configural invariance multi-group measurement model.

Following this procedure, the statistical significance was calculated in Table 5.56 for each separate model that was estimated for each non-invariant item, until partial metric equivalence was established.

Upon fitting each multi-group partial weak invariance model per non-invariant stress item, the degrees of freedom were examined. The degrees of freedom tallied with the calculations[27], which implied that the models were specified and estimated correctly.

[27] The degrees of freedom calculations for each partial weak invariance model were calculated separately by the researcher in order to ensure each model was specified correctly.

Table 5.56

Statistical significance of the scaled chi-squared difference statistic: a test of partial metric equivalence (items)

HYPOTHESIS	SATORRA-BENTLER CHI SQUARE	NORMAL THEORY CHI-SQUARE	DF	Cd	SCALED DIFFERENCE IN S-B CHI-SQUARE	PROB S-B CHI-SQUARE DIFF	PROB SCALED S-B CHI-SQUARE DIFF
H_a CONFIGURAL INVARIANCE MODEL	7412.868	8900.462	3232				
PARTIAL WEAK INVARIANCE H_{03J} MODEL	7540.183	9021.832	3281				
PARTIAL WEAK INVARIANCE H_{03k} MODEL	7532.177	9011.853	3280				
PARTIAL WEAK INVARIANCE H_{03l} MODEL	7517.306	8994.999	3279				
DIFF (H_{03j}-H_a) PARTIAL METRIC EQIVALENCE	**127.315**		**49**	**0.920999**	**131.7808609**	**6.89062E-09**	**1.63084E-09**
DIFF (H_{03k}-H_a) PARTIAL METRIC EQIVALENCE	**119.309**		**48**	**0.911632**	**122.1885791**	**5.30133E-08**	**2.114494-08**
DIFF (H_{03l}-H_a) PARTIAL METRIC EQIVALENCE	**104.438**		**47**	**0.914281**	**103.4004145**	**3.00866E-06**	**4.06448E-08**

H_{03j}-H_a = Item LA35 allowed to vary; H_{03k}-H_a = Items LA35 and JS27 allowed to vary; and H_{03l}-H_a = Items LA35, JS27 and RA6 allowed to vary.

If the probability of observing the scaled chi-squared difference in a multi-group sample under the null hypothesis of no difference in fit between the partial weak and configural invariance models in the parameter (H_0:$\chi^2 = 0$) is smaller than or equal to .05, the null hypothesis is rejected in favour of the hypothesis that the fit of the multi-group configural invariance and multi-group partial weak invariance models differ in the parameter. As can be seen in Table 5.56, the null hypothesis of no difference in model fit in the parameter was rejected ($p < .05$) for all three partial invariance models. This implies a lack of equivalence of the factor loadings across the two gender groups even when the

non-invariant factor loadings are allowed to vary (i.e. lack of partial metric equivalence) when using statistical significance as the yardstick of equivalence.

The results for the test of practical significance of the difference in multi-group measurement model fit are shown in Table 5.57.

Table 5.57
Practical significance of the CFI, Gamma Hat and MacDonald difference statistic: a test of partial metric equivalence per item

MODEL	CFI	Γ_1	Mc
CONFIGURAL INVARIANCE H_a MODEL	0.971	0.9283444	0.10259152
PARTIAL WEAK INVARIANCE H_{03j} MODEL	0.971	0.927089881	0.098273586
PARTIAL WEAK INVARIANCE H_{03k} MODEL	0.971	0.927206437	0.098667467
PARTIAL WEAK INVARIANCE H_{03l} MODEL	0.971	0.927425059	0.099410255
DIFF [H_{03j}-H_a; TEST OF PARTIAL METRIC EQUIVALENCE]	**0**	**-0.001254519**	**-0.004317934**
DIFF [H_{03k}-H_a; TEST OF PARTIAL METRIC EQUIVALENCE]	**0**	**-0.001137963**	**-0.003924053**
DIFF [H_{03l}-H_a; TEST OF PARTIAL METRIC EQUIVALENCE]	**0**	**-0.000919341**	**-0.003181265**

H_{03j}-H_a = Item LA35 allowed to vary; H_{03k}-H_a = Items LA35 and JS27 allowed to vary; H_{03l}-H_a = Items LA35, JS27 and RA6 allowed to vary..

Partial metric equivalence may be concluded if the partial weak invariance multi-group measurement model in which non-invariant factor loadings are allowed to vary does not fit practically significantly poorer than the configural invariance multi-group measurement model. As indicated in Table 5.57, the partial weak invariance multi-group measurement model in which the non-invariant slopes of the regression of the indicator variables LA35, JS27 and RA6 were not constrained to be equal across gender groups, the change in CFI, Γ_1 and Mc was less than the critical thresholds of -0.01, -0.001 and -0.02 respectively. This provided support for partial metric equivalence when using practical significance as the yardstick of equivalence.

5.12.1 DECISION ON THE RESULTS OF PARTIAL METRIC EQUIVALENCE

Partial metric equivalence was adequately supported when a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, and in which all parameters were estimated freely across the gender samples, but for the slopes of the regression of 56 indicator variables on the latent variables, did not fit practically significantly poorer than a multi-group measurement model in which the structure of the model is constrained to be the same across gender groups but all parameters are estimated freely. When the partial weak invariance model which allowed for the factor loadings of the non-invariant items LA35, JS27 and

RA6 to vary across gender groups was compared to the configural invariance model, partial metric equivalence was displayed, as the results for all three practical fit indices were below the critical thresholds. Cheung and Rensvold (1999) have argued that a comparison of factor means is still feasible when most of the items are invariant, and that under these conditions, failure to achieve full factorial invariance is trivial from a practical point of view. It is therefore concluded that with 56 of the 59 SWSI items shown to be invariant, adequate support for partial metric equivalence has been displayed. Due to only a small portion of the model constituting of non-invariant factor loadings, it can be assumed these items will not significantly affect cross-group comparisons (Cheung & Rensvold, 1999). This finding implies a partial lack of non-uniform item bias, when the three non-invariant items are acknowledged.

Under these conditions, strong invariance was tested, however, in doing so, the difference in the factor loadings of the items LA35, JS27 and RA6 was formally acknowledged (in the model specification) in the multi-group strong invariance measurement model.

5.13 STRONG INVARIANCE

Upon finding acceptable model fit for the weak invariance model, as well as establishing partial metric equivalence, strong invariance[28] was tested. This test aimed to establish whether a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, and in which all parameters were estimated freely across the samples, but for the 56 invariant factor loadings and the vector of regression intercepts, demonstrated acceptable fit when fitted to the samples simultaneously in a multi-group analysis.

Operational hypothesis 4 was tested by testing H_{04}: RMSEA ≤ .05[29].

The SWSI model, in which the structure of the model and the factor loadings were constrained to be the same across gender groups (except for the factor loadings of items LA35, JS27 and RA6 which are allowed to vary), and in which the vector of the regression intercepts were constrained to be the same across gender groups was fitted to the male (*N*=460) and female (*N*=460) samples. A visual representation of the fitted multi-group strong invariance SWSI measurement model is shown in Figure 5.9.

[28] A refinement in the taxonomy seems to be called for to make provision for partial invariance and equivalence. The number of combinations, however, seems rather daunting. On each current form of invariance (excluding configural invariance) and equivalence, two possible outcomes can be found. This implies 32 different possible combinations of forms of partial invariance and partial equivalence. The magnitude of the challenge has dissuaded attempts to propose such a refined taxonomy in this study.

[29] Strictly speaking operational hypothesis 4 and H_{04} and the hypotheses tested here are not exactly the same.

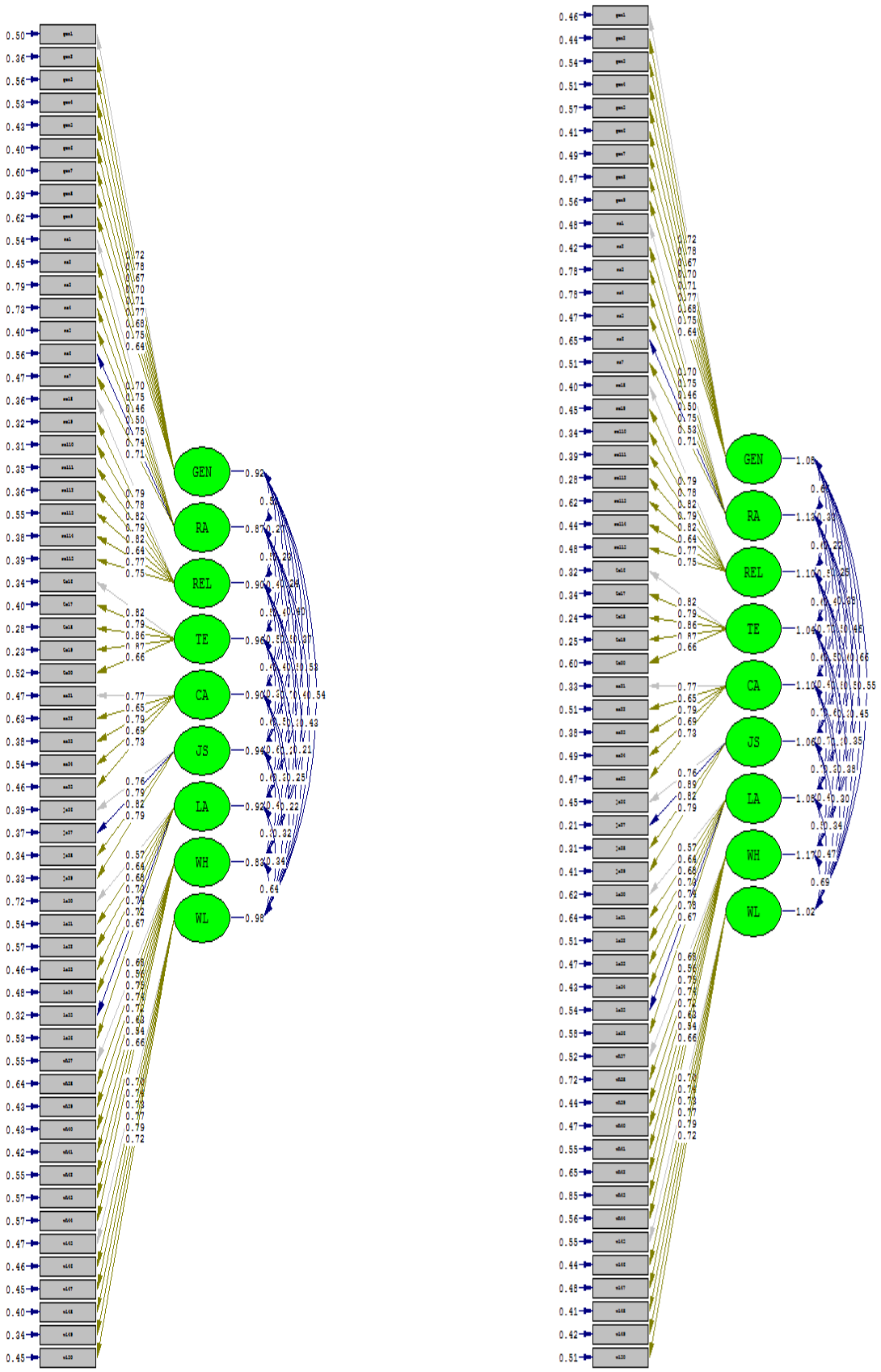

Figure 5.9. Representation of the fitted multi-group SWSI strong invariance measurement model for the male and female sample respectively

5.13.1 MEASUREMENT MODEL FIT INDICES

The strong invariance measurement model converged in 36 iterations. The spectrum of fit statistics is shown in Table 5.58.

Table 5.58

Goodness of fit statistics for the SWSI strong invariance measurement model

Degrees of Freedom = 3338
Minimum Fit Function Chi-Square = 8561.624 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 9283.135 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 7848.710 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 4510.710
90 Percent Confidence Interval for NCP = (4255.741 ; 4773.171)

Minimum Fit Function Value = 9.326
Population Discrepancy Function Value (F0) = 4.914
90 Percent Confidence Interval for F0 = (4.636 ; 5.200)
Root Mean Square Error of Approximation (RMSEA) = 0.0543
90 Percent Confidence Interval for RMSEA = (0.0527 ; 0.0558)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000

Expected Cross-Validation Index (ECVI) = 9.247
90 Percent Confidence Interval for ECVI = (8.841 ; 9.404)
ECVI for Saturated Model = 3.856
ECVI for Independence Model = 161.365

Chi-Square for Independence Model with 3422 Degrees of Freedom = 148015.476
Independence AIC = 148251.476
Model AIC = 8488.710
Saturated AIC = 7080.000
Independence CAIC = 148938.752
Model CAIC = 10352.509
Saturated CAIC = 27698.283

Normed Fit Index (NFI) = 0.947
Non-Normed Fit Index (NNFI) = 0.968
Parsimony Normed Fit Index (PNFI) = 0.924
Comparative Fit Index (CFI) = 0.969
Incremental Fit Index (IFI) = 0.969
Relative Fit Index (RFI) = 0.946

Critical N (CN) = 413.994

Upon first inspection of Table 5.58, the degrees of freedom were indicated to be 3338. This did not correspond with the calculations in Table 4.1, as the freed non-invariant factor loadings were not foreseen and taken into account in the calculations reflected in Table 4.1. A recalculation resulted in the degrees of freedom for this multi-group measurement invariance model, which takes the three

non-invariant factor loadings into account, to be 3658-320=3338. This, therefore, provided evidence that the model was specified correctly.

Due to not testing the exact fit null hypothesis, the following close fit null hypothesis, which explicitly assumes that the measurement model only approximates the processes that operated in reality to create the observed covariance matrix (Browne & Cudeck, 1993), was tested:

$$H_{04}: \text{RMSEA} \leq .05$$

$$H_{a4}: \text{RMSEA} > .05$$

A RMSEA value of .0543 was obtained (Table 5.58), indicating that the measurement model showed reasonable to good fit in the sample. The confidence interval for the RMSEA was (.0527 ; .0558). The small interval indicated a higher level of precision in reflecting the model fit in the population (Byrne, 2001). The p-value for Test of Close Fit (H_{04}: RMSEA < .05) was 1.00, therefore the close fit null hypothesis H_{04}: RMSEA ≤ .05 was not rejected (p > .05). The position that the measurement model thus showed close fit was therefore a tenable position. The measurement model was therefore a plausible explanation and the model approximately reproduced the observed covariance matrix, but not perfectly.

5.13.2 DECISION ON THE SUCCESS OF STRONG INVARIANCE

Upon fitting the strong invariance measurement model (which allows for three freed non-invariant factor loadings), it was established that the multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, and in which all parameters were estimated freely across the gender samples, but for the 56 factor loadings and the vector of regression intercepts that were constrained to be equal, demonstrated acceptable fit when fitted to the gender samples simultaneously in a multi-group analysis.

Strong invariance was displayed, which indicated that the 56 regression slopes and all the intercepts were the same across gender samples because the observed covariance matrices of the two samples were adequately explained. This finding implies that it is an acceptable position to hold that the intercepts of the items on the latent variable they represent are the same across gender samples. Support is thus provided for the position that the items operate in approximately the same way across the gender samples in the way that they reflect the underlying latent variables they are meant to reflect. Together, these results suggest a lack of uniform item bias. It can as yet, however, not be claimed that a model where the intercepts of the regression of the items on the latent variables they represent are not the same across the samples, does not fit the data better. To

investigate this, scalar equivalence was tested. In addition it needs to be acknowledged that three of the items in the SWSI display non-uniform item bias. With regards to these items, the question as to whether the intercept parameters also differ across groups was not really of any practical relevance, since it was already known that the items were biased.

5.14 SCALAR EQUIVALENCE

A finding of strong invariance allowed for the testing of scalar equivalence. This test of scalar equivalence via multi-group confirmatory factor analysis using LISREL determines whether the multi-group (strong invariance) measurement model in which the 56 parameters in $\boldsymbol{\Lambda}^{x}$ and all parameters in $\boldsymbol{\tau}$ are constrained to be equal across groups fits (a) statistically significantly better or (b) practically significantly better than a multi-group (configural invariance) measurement model in which all model parameters are estimated freely. The decision on whether the multi-group SWSI strong invariance measurement model shows scalar equivalence is based on the question whether the multi-group strong invariance model fits practically significantly poorer than the multi-group configural invariance model. The results of the test whether the multi-group strong invariance model fits statistically significantly poorer than the multi-group configural invariance model are nonetheless also provided.

The results of the test of the statistical significance of the difference in multi-group measurement model fit for this analysis, is shown in Table 5.59.

Table 5.59
Statistical significance of the scaled chi-squared difference statistic: a test of scalar equivalence

HYPOTHESIS	SATORRA-BENTLER CHI SQUARE	NORMAL THEORY CHI-SQUARE	DF	Cd	SCALED DIFFERENCE IN S-B CHI-SQUARE	PROB S-B CHI-SQUARE DIFF	PROB SCALED S-B CHI-SQUARE DIFF
H_a CONFIGURAL INVARIANCE MODEL	7412.868	8900.462	3232				
STRONG INVARIANCE H_{04} MODEL	7848.710	9283.135	3338				
DIFF (H_{04}-H_a) SCALAR EQUIVALENCE	**435.842**		**106**	**0.636432**	**601.278938**	**1.44746E-41**	**2.93626E-70**

If the probability of observing the scaled chi-squared difference in a multi-group sample under the null hypothesis of no difference in fit between the strong and configural invariance models in the parameter (H_0:$\chi^2 = 0$) is smaller than or equal to .05, the null hypothesis is rejected in favour of the hypothesis that the fit of the multi-group configural invariance and multi-group strong invariance models differ in the parameter. As can be seen in Table 5.59, the null hypothesis of no difference in model fit in the parameter was rejected ($p \leq .05$), and this implies a lack of equivalence of intercepts across the two gender groups (i.e. lack of scalar equivalence) when using statistical significance as the yardstick of equivalence.

The results for the tests of practical significance of the difference in multi-group measurement model fit for this analysis are shown in Table 5.60.

Table 5.60

Practical significance of the CFI, Gamma Hat and MacDonald difference statistic: a test of scalar equivalence

MODEL	CFI	Γ_1	Mc
CONFIGURAL INVARIANCE H_a MODEL	0.971	0.9283444	0.10259152
STRONG INVARIANCE H_{04} MODEL	0.969	0.923115436	0.085691641
DIFF [H_{04}-H_a; TEST OF SCALAR EQUIVALENCE]	**-0.002**	**-0.005228964**	**-0.017**

CFI= Comparative Fit Index; Γ_1= Gamma Hat fit index; Mc= McDonald Non-centrality index

Scalar equivalence may be concluded if the strong invariance multi-group measurement model (which took the non-invariant factor loadings into account) does not fit practically significantly poorer than the configural invariance multi-group measurement model. As indicated in Table 5.60, the change in CFI and Mc was less than the critical thresholds of -0.01 and -0.02 respectively, therefore the difference in fit is considered practically insignificant given results of these two indices. However, the Gamma Hat fit index change was greater than the critical threshold of -0.001, and therefore the difference in model fit was considered practically significant. In terms of the decision-rule specified in Chapter 4, scalar equivalence cannot be concluded as all three criteria were not met. It was therefore concluded that a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, and in which all parameters were estimated freely across the gender samples, but for 56 factor loadings and all intercepts that were constrained to be equal, fitted practically significantly poorer than a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups but

all parameters were estimated freely. This, in turn, suggests that the intercept parameter estimates for one or more items differ practically significantly across the two gender groups.

5.14.1 DECISION ON THE RESULTS OF STRONG INVARIANCE AND SCALAR EQUIVALENCE

Upon fitting the strong invariance multi-group measurement model, the results indicated acceptable model fit, thereby permitting the conclusion that a lack of uniform item bias was tenable with regards to the SWISI. When comparing the strong invariance multi-group measurement model to the configural invariance multi-group measurement model (i.e., one with fewer constraints), scalar equivalence was not supported, as the difference in the Gamma Hat fit index was greater than the critical threshold of -0.001. This implied that although strong invariance is tenable in that the multi-group model in which the intercepts were constrained to be equal across groups could closely reproduce the observed covariance matrices, the configural invariance model in which the intercepts are allowed to vary could reproduce the observed covariance matrices practically significantly better. This suggests that a lack of equivalence of regression intercepts exists across the gender groups on one or more items was therefore a more convincing position.

As with metric equivalence, inadequate support for scalar equivalence required additional tests to determine the source of non-invariance with regards to the regression intercepts. As scalar equivalence, or at least partial scalar equivalence, is a necessary prerequisite for cross-group comparisons of observed means and for cross-group comparisons of latent means, accurate identification of non-invariant intercepts is necessary.

5.15 PARTIAL SCALAR EQUIVALENCE

A lack of support for scalar equivalence resulted in the need for additional tests to determine the source of non-equivalence, resulting in a finding of partial scalar equivalence. The non-invariant factor loadings have already been identified (LA35, JS27, and RA6), and now the non-invariant intercepts must also be correctly identified (Cheung & Rensvold, 1999).

The procedure that was followed involved (a) examining the intercepts ($\boldsymbol{\tau}$) of the unconstrained configural invariance model to identify the greatest difference between gender groups, and (b) testing for partial scalar equivalence when lifting the intercept equality constraint one item at a time, starting with the item with the most dissimilar intercepts across the two gender samples .

All intercepts were considered for possible non-invariance. The differences between the male and female tau estimates in the configural invariance output were calculated from the unstandardised

solution[30] and the absolute values rank ordered from high to low in Table 5.61 in order to identify the items with the most dissimilar intercepts.

Table 5.61

Tau_X_difference

	Configural invariance: LISREL Estimate Tau_X		Tau_X_diff	Rank order from largest to smallest	
	Male	**Female**	**Difference**	**Item**	**Difference**
GEN1	2.389	2.572	-0.183	**WH40**	**0.363**
GEN2	1.946	2.239	-0.293	**GEN2**	**0.293**
GEN3	1.898	2.172	-0.274	**WL47**	**0.277**
GEN4	2.202	2.350	-0.148	**GEN3**	**0.274**
GEN5	2.085	2.148	-0.063	**WH39**	**0.263**
GEN6	2.011	2.091	-0.080	**WH42**	**0.261**
GEN7	2.500	2.561	-0.061	**WL48**	**0.243**
GEN8	1.728	1.926	-0.198	**CA22**	**0.235**
GEN9	2.039	2.150	-0.111	**WH41**	**0.217**
RA1	1.946	2.137	-0.191	**WH44**	**0.213**
RA2	2.002	2.165	-0.163	**GEN8**	**0.198**
RA3	2.422	2.372	0.050	**TE20**	**0.193**
RA4	2.104	2.080	0.024	**RA1**	**0.191**
RA5	1.907	2.015	-0.108	**TE17**	**0.191**
RA6	2.307	2.159	0.148	**WH37**	**0.189**
RA7	2.093	2.157	-0.064	**GEN1**	**0.183**
REL8	2.224	2.154	0.070	**WL45**	**0.169**
REL9	1.978	1.978	0.000	**CA21**	**0.168**
REL10	2.343	2.337	0.006	**CA23**	**0.165**
REL11	1.967	1.843	0.124	**RA2**	**0.163**
REL12	1.802	1.837	-0.035	**WH38**	**0.161**
REL13	1.783	1.907	-0.124	**LA36**	**0.152**
REL14	2.054	2.063	-0.009	**RA6**	**0.148**
REL15	1.722	1.839	-0.117	**GEN4**	**0.148**
TE16	2.111	2.011	0.100	**LA32**	**0.141**
TE17	2.239	2.048	0.191	**REL11**	**0.124**
TE18	2.354	2.257	0.097	**REL13**	**0.124**
TE19	2.267	2.207	0.060	**WL49**	**0.120**
TE20	2.246	2.439	-0.193	**REL15**	**0.117**
CA21	2.711	2.543	0.168	**GEN9**	**0.111**
CA22	2.728	2.493	0.235	**RA5**	**0.108**
CA23	2.576	2.411	0.165	LA30	0.104
CA24	2.170	2.115	0.055	TE16	0.100
CA25	2.639	2.602	0.037	JS29	0.100
JS26	2.326	2.411	-0.085	JS28	0.097
JS27	2.600	2.663	-0.063	TE18	0.097
JS28	2.446	2.543	-0.097	JS26	0.085
JS29	2.148	2.248	-0.100	WL46	0.084

[30] In the completely standardised solution all intercepts would be zero.

LA30	2.378	2.274	0.104	GEN6	0.080
LA31	2.239	2.198	0.041	LA35	0.074
LA32	2.215	2.074	0.141	WL50	0.072
LA33	2.196	2.157	0.039	REL8	0.070
LA34	2.472	2.520	-0.048	WH43	0.065
LA35	2.604	2.678	-0.074	RA7	0.064
LA36	2.352	2.504	-0.152	GEN5	0.063
WH37	2.504	2.693	-0.189	JS27	0.063
WH38	1.891	2.052	-0.161	GEN7	0.061
WH39	2.309	2.572	-0.263	TE19	0.060
WH40	1.891	2.254	-0.363	CA24	0.055
WH41	1.657	1.874	-0.217	RA3	0.050
WH42	1.717	1.978	-0.261	LA34	0.048
WH43	1.398	1.463	-0.065	LA31	0.041
WH44	2.180	2.393	-0.213	LA33	0.039
WL45	2.483	2.652	-0.169	CA25	0.037
WL46	2.193	2.109	0.084	REL12	0.035
WL47	2.470	2.193	0.277	RA4	0.024
WL48	2.213	1.970	0.243	REL14	0.009
WL49	2.489	2.609	-0.120	REL10	0.006
WL50	2.278	2.350	-0.072	REL9	0.000

GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface. **Items with non-invariant intercepts are in boldface.**

The process to identify non-invariant intercepts continued with the performance of a series of further tests. Using the rank ordered absolute values, in which the items with the greatest difference in τ between gender groups was identified, a multi-group model was estimated in which all intercepts were constrained to be equal, but for the item that had the most dissimilar intercept in the fully constrained solution, while still honouring the results of the tests for partial metric equivalence. Therefore, under partial strong invariance, the 56 invariant factor loadings were constrained to be equal across gender groups but the three non-invariant factor loadings that were allowed to vary, and the invariant intercepts were constrained to be equal across gender groups, while the most dissimilar intercept was permitted to vary. Under a finding of a lack of partial scalar equivalence, the process was repeated by subsequently freeing the item that obtained the second most dissimilar intercept in the fully unconstrained solution. This process continued until practical insignificance was achieved, indicating that a partial strong invariance multi-group measurement model has been found in which the non-invariant factor loadings and intercepts are allowed to vary, and that did not fit practically significantly poorer than the configural invariance multi-group measurement model.

Following this procedure, the statistical significance of the scaled Satorra-Bentler difference statistic (Table 5.62) was calculated for each of the partial strong invariance multi-group models in which the

most dissimilar intercepts were allowed to vary across groups, until practical significance was established.

Upon fitting each multi-group partial strong invariance model per non-invariant stress item, the resultant degrees of freedom of the specified model were examined. The degrees of freedom from the results tallied up with the calculations[31], confirming that the models were estimated correctly.

Table 5.62
Statistical significance of the scaled chi-squared difference statistic: a test of partial scalar equivalence per intercept

HYPOTHESIS	SATORRA-BENTLER CHI SQUARE	NORMAL THEORY CHI-SQUARE	DF	Cd	SCALED DIFFERENCE IN S-B CHI-SQUARE	PROB S-B CHI-SQUARE DIFF	PROB SCALED S-B CHI-SQUARE DIFF
Ha CONFIGURAL INVARIANCE MODEL	7412.868	8900.462	3232				
PARTIAL STRONG INVARIANCE H_{04a} MODEL	7832.189	9267.142	3337				
PARTIAL STRONG INVARIANCE H_{04b} MODEL	7811.291	9246.507	3336				
PARTIAL STRONG INVARIANCE H_{04c} MODEL	7750.515	9183.38	3335				
PARTIAL STRONG INVARIANCE H_{04d} MODEL	7730.347	9163.217	3334				
PARTIAL STRONG INVARIANCE H_{04e} MODEL	7727.099	9160.141	3333				
PARTIAL STRONG INVARIANCE H_{04f} MODEL	7715.667	9148.917	3332				
PARTIAL STRONG INVARIANCE H_{04g} MODEL	7676.036	9114.262	3331				

[31] The degrees of freedom calculations for each partial strong invariance model were calculated separately by the researcher in order to ensure each model was specified correctly.

PARTIAL STRONG INVARIANCE H_{04h} MODEL	7673.215	9112.817	3330
PARTIAL STRONG INVARIANCE H_{04i} MODEL	7671.654	9112.476	3329
PARTIAL STRONG INVRIANCE H_{04j} MODEL	7670.448	9112.059	3328
PARTIAL STRONG INVARIANCE H_{04k} MODEL	7667.302	9110.273	3327
PARTIAL STRONG INVARIANCE H04l MODEL	7650.148	9096.008	3326
PARTIAL STRONG INVARIANCE $H04_m$ MODEL	7635.348	9081.009	3325
PARTIAL STRONG INVARIANCE H_{04n} MODEL	7629.935	9076.663	3324
PARTIAL STRONG INVARIANCE H_{04o} MODEL	7628.227	9075.551	3323
PARTIAL STORNG INVARIANCE H_{04p} MODEL	7623.454	9071.62	3322
PARTIAL STRONG INVARIANCE H_{04q} MODEL	7620.317	9069.545	3321
PARTIAL STRONG INVARIANCE H_{04r} MODEL	7615.105	9064.679	3320
PARTIAL STRONG INVARIANCE H_{04s} MODEL	7608.976	9059.409	3319
PARTIAL STRONG INVARIANCE H_{04t} MODEL	7602.028	9053.714	3318
PARTIAL STRONG INVARIANCE H_{04u} MODEL	7599.177	9051.472	3317

PARTIAL STRONG INVARIANCE H_{04v} MODEL	7590.872	9044.239	3316				
PARTIAL STRONG INVARIANCE H_{04w} MODEL	7585.199	9041.093	3315				
PARTIAL STRONG INVARIANCE H_{04x} MODEL	7585.521	9042.373	3314				
PARTIAL STRONG INVARIANCE H_{04y} MODEL	7577.812	9035.624	3313				
PARTIAL STRONG INVARIANCE H_{04z} MODEL	7569.553	9028.852	3312				
PARTIAL STRONG INVARIANCE H_{04aa} MODEL	7562.022	9021.884	3311				
PARTIAL STRONG INVARIANCE H_{04bb} MODEL	7559.913	9020.726	3310				
PARTIAL STRONG INVARIANCE H_{04cc} MODEL	7552.761	9014.598	3309				
PARTIAL STRONG INVARIANCE H_{04dd} MODEL	7551.852	9014.267	3308				
PARTIAL STRONG INVARIANCE H_{04ee} MODEL	7548.548	9011.769	3307				
DIFF (H_{04a}-H_a) PARTIAL SCALAR EQUIVALENCE	**419.321**		**105**	**0.645623**	**567.9476398**	**3.78133E-39**	**1.1336E-64**
DIFF (H_{04b}-H_a) PARTIAL SCALAR EQUIVALENCE	**398.423**		**104**	**0.657257**	**526.4991316**	**4.84593E-36**	**1.03112E-57**
DIFF (H_{04c}-H_a) PARTIAL SCALAR EQUIVALENCE	**337.647**		**103**	**0.688974**	**410.6367771**	**9.62245E-27**	**2.48241E-38**
DIFF (H_{04d}-H_a) PARTIAL SCALAR EQUIVALENCE	**317.479**		**102**	**0.699899**	**375.418673**	**5.9429E-24**	**6.360499E-33**
DIFF (H_{04e}-H_a) PARTIAL SCALAR EQUIVALENCE	**314.231**		**101**	**0.698398**	**371.8210731**	**1.01972E-23**	**1.23438E-32**

DIFF (H_{04f}-H_a) PARTIAL SCALAR EQUIVALENCE	302.799	100	0.703581	353.1291052	2.87001E-22	5.93E-30
DIFF (H_{04g}-H_a) PARTIAL SCALAR EQUIVALENCE	263.168	99	0.752791	284.0097265	7.83427E-17	9.0369E-20
DIFF (H_{04h}-Ha) PARTIAL SCALAR EQUIVALENCE	260.347	98	0.756791	280.5993728	1.16511E-16	1.6328E-19
DIFF (H_{04i}-H_a) PARTIAL SCALAR EQUIVALENCE	258.786	97	0.759117	279.2902331	1.16051E-16	1.4712E-19
DIFF (H_{04j}-H_a) PARTIAL SCALAR EQUIVALENCE	257.58	96	0.759241	278.6953862	1.031E-16	1.0434E-19
DIFF (H_{04k}-H_a) PARTIAL SCALAR EQUIVALENCE	254.434	95	0.763641	274.7508081	1.70157E-16	2.2544E-19
DIFF (H_{04l}-H_a) PARTIAL SCALAR EQUIVALENCE	237.28	94	0.787418	248.3382158	2.28458E-14	7.1769E-16
DIFF (H_{04m}-H_a) PARTIAL SCALAR EQUIVALENCE	222.48	93	0.795266	227.0271883	1.28837E-12	3.2865E-13
DIFF (H_{04n}-H_a) PARTIAL SCALAR EQUIVALENCE	217.067	92	0.800888	220.0069292	4.12135E-12	1.7205E-12
DIFF (H_{04o}-H_a) PARTIAL SCALAR EQUIVALENCE	215.359	91	0.80102	218.58253	4.38819E-12	1.6794E-12
DIFF (H_{04p}-H_a) PARTIAL SCALAR EQUIVALENCE	210.586	90	0.805163	212.575711	1.1561E-11	6.4251E-12
DIFF (H_{04q}-H_a) PARTIAL SCALAR EQUIVALENCE	207.449	89	0.808957	209.0135072	1.8768E-11	1.1842E-11
DIFF (H_{04r}-H_a) PARTIAL SCALAR EQUIVALENCE	202.237	88	0.81125	202.4246178	5.58059E-11	5.2841E-11
DIFF (H_{04s}-H_a) PARTIAL SCALAR EQUIVALENCE	196.108	87	0.817049	194.5379331	2.1427E-10	33579E-10
DIFF (H_{O4t}-H_a) PARTIAL SCALAR EQUIVALENCE	189.16	86	0.825786	185.5832171	1.02128E-09	2.7794E-09
DIFF (H_{04u}-H_a) PARTIAL SCALAR EQUIVALENCE	186.309	85	0.827413	182.5086522	1.51537E-09	4.377E-09
DIFF (H_{04v}-H_a) PARTIAL SCALAR EQUIVALENCE	178.004	84	0.836912	171.7946197	1.02065E-08	5.4658E-08
DIFF (H_{04w}-Ha) PARTIAL SCALAR EQUIVALENCE	172.331	83	0.851666	165.124686	3.24395E-08	2.1899E-07
DIFF (H_{04x}-H_a) PARTIAL SCALAR	172.653	82	0.852291	166.5054059	2.02232E-08	1.0568E-07

EQUIVALENCE						
DIFF (H_{04y}-H_a) PARTIAL SCALAR EQUIVALENCE	**164.944**	**81**	**0.861269**	**156.9335555**	**1.10241E-07**	**8.8499E-07**
DIFF (H_{04z}-H_a) PARTIAL SCALAR EQUIVALENCE	**156.685**	**80**	**0.873953**	**146.9072452**	**6.63496E-07**	**7.5955E-06**
DIFF (H_{04aa}-Ha) PARTIAL SCALAR EQUIVALENCE	**149.154**	**79**	**0.881084**	**137.8097799**	**3.14879E-06**	**4.7268E-05**
DIFF (H_{04bb}-Ha) PARTIAL SCALAR EQUIVALENCE	**147.045**	**78**	**0.884708**	**135.9363577**	**3.79562E-06**	**5.378E-05**
DIFF (H_{04cc}-H_a) PARTIAL SCALAR EQUIVALENCE	**139.893**	**77**	**0.894391**	**127.613085**	**1.56365E-05**	**0.00025667**
DIFF (H_{04dd}-H_a) PARTIAL SCALAR EQUIVALENCE	**138.984**	**76**	**0.8948**	**127.1848207**	**1.41129E-05**	**0.00021323**
DIFF (H_{04ee}-H_a) PARTIAL SCALAR EQUIVALENCE	**135.68**	**75**	**0.899261**	**123.7760815**	**2.26351E-05**	**0.00033823**

H_{04a}-H_a = Item WH40 allowed to vary; H_{04b}-H_a = Items WH40 and GEN2 allowed to vary; H_{04c}-H_a = Items WH40, GEN2 and WL47 allowed to vary, H_{04d}-H_a = Items WH40, GEN2, WL47 and GEN3 allowed to vary; H_{04e}-H_a = Items WH40, GEN2, WL47, GEN3 and WH39 allowed to vary; H_{04f}-H_a = Items WH40, GEN2, WL47, GEN3, WH39 and WH42 allowed to vary; H_{04g}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42 and WL48 allowed to vary; H_{04h}-H_a = Items WH40, GEN2, WL47 GEN3, WH39, WH42, WL48, and CA22 allowed to vary; H_{04i}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, and WH41 allowed to vary; H_{04j}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, and WH44 allowed to vary; H_{04k}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, and GEN8 allowed to vary; H_{04l}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, and TE20 allowed to vary; H_{04m}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, and RA1 allowed to vary; H_{04n}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20,RA1, and TE17 allowed to vary; H_{04o}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, and WH37 allowed to vary; H_{04p}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, and GEN1 allowed to vary; H_{04q}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, and WL45 allowed to vary; H_{04r}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WH45, and CA21 allowed to vary; H_{04s}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, and CA23 allowed to vary; H_{04t}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, and RA2 allowed to vary; H_{04u}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2 and WH38 allowed to vary; H_{04v}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38 and LA36 allowed to vary; H_{04w}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36 and RA6 allowed to vary; H_{04x}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36 and GEN4 allowed to vary; H_{04y}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4 and LA32 allowed to vary; H_{04z}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, and REL11 allowed to vary; H_{04aa}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11 and REL13 allowed to vary; H_{04bb}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11, REL13 and WL49 allowed to vary; H_{04cc}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11, REL13, WL49 and REL15 allowed to vary; H_{04dd}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11, REL13, WL49, REL15 and GEN9 allowed to vary; H_{04ee}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11, REL13, WL49, REL15, GEN9 and RA5 allowed to vary.

If the probability of observing the scaled chi-squared difference in a multi-group sample under the null hypothesis of no difference in fit between the partial strong and configural invariance models in the parameter (H_0: $\chi^2 = 0$) is smaller than or equal to .05, the null hypothesis is rejected in favour of the hypothesis that that the fit of the partial strong and configural invariance models are different in the parameter. As is evident from Table 5.62, the null hypothesis of no difference in model fit was rejected ($p \leq .05$) for each of the partial strong invariance models that were compared to the configural invariance model. This implied a lack of equivalence of regression intercepts across the two gender groups even when the indicated non-invariant intercepts were allowed to vary (i.e. lack of partial scalar equivalence) when using statistical significance as the yardstick of equivalence.

The results for the test of practical significance of the difference in multi-group measurement model fit, is shown in Table 5.63.

Table 5.63

Practical significance of the CFI, Gamma Hat and MacDonald difference statistic: a test of partial scalar equivalence per item intercept

MODEL	**CFI**	Γ_1	**Mc**
CONFIGURAL INVARIANCE H_a MODEL	0.971	0.9283444	0.10259152
PARTIAL STRONG INVARIANCE H_{04a} MODEL	0.969	0.923361034	0.086423124
PARTIAL STRONG INVARIANCE H_{04b} MODEL	0.969	0.923679061	0.087379026
PARTIAL STRONG INVARIANCE H_{04c} MODEL	0.969	0.924619966	0.090265496
PARTIAL STRONG INVARIANCE H04d MODEL	0.97	0.924924360	0.091218277
PARTIAL STRONG INVARIANCE H_{04e} MODEL	0.97	0.924953360	0.091309541
PARTIAL STRONG INVARIANCE H_{04f} MODEL	0.97	0.925127401	0.091859045
PARTIAL STRONG INVARIANCE H_{04g} MODEL	0.97	0.925737059	0.093808482
PARTIAL STRONG INVARIANCE H_{04h}- MODEL	0.97	0.925766111	0.093902337
PARTIAL STRONG INVARIANCE H_{04i} MODEL	0.97	0.925766111	0.093902337
PARTIAL STRONG INVRIANCE H_{04j} MODEL	0.97	0.925780637	0.093949300
PARTIAL STRONG INVARIANCE H_{04k} MODEL	0.97	0.925809691	0.094043297
PARTIAL STRONG INVARIANCE H_{04l} MODEL	0.97	0.926071260	0.094893507
PARTIAL STRONG INVARIANCE H_{04m} MODEL	0.97	0.926289348	0.095607883
PARTIAL STRONG INVARIANCE H_{04n} MODEL	0.97	0.926347522	0.095799291
PARTIAL STRONG INVARIANCE H_{04o} MODEL	0.97	0.926362066	0.095847202
PARTIAL STRONG INVARIANCE H_{04p} MODEL	0.97	0.926420249	0.096039088
PARTIAL STRONG INVARIANCE H_{04q} MODEL	0.97	0.926463891	0.096183255
PARTIAL STRONG INVARIANCE H_{04r} MODEL	0.97	0.926522087	0.096375814
PARTIAL STRONG INVARIANCE H_{04s} MODEL	0.97	0.926609395	0.096665376
PARTIAL STRONG INVARIANCE H_{04t} MODEL	0.97	0.926696719	0.096955807
PARTIAL STRONG INVARIANCE H_{04u} MODEL	0.97	0.926725831	0.097053812
PARTIAL STRONG INVARIANCE H_{04v} MODEL	0.97	0.926842295	0.097441800

PARTIAL STRONG INVARIANCE H_{04w} MODEL	0.97	0.926915101	0.097685709
PARTIAL STRONG INVARIANCE H_{04x} MODEL	0.97	0.926900539	0.097636879
PARTIAL STRONG INVARIANCE H_{04y} MODEL	0.971	0.927002482	0.097979207
PARTIAL STRONG INVARIANCE H_{04z} MODEL	0.971	0.927119017	0.098371908
PARTIAL STRONG INVARIANCE H_{04aa} MODEL	0.971	0.927221009	0.098716813
PARTIAL STRONG INVARIANCE H_{04bb} MODEL	0.971	0.927235581	0.098766184
PARTIAL STRONG INVARIANCE H_{04cc} MODEL	0.971	0.927337598	0.099112471
PARTIAL STRONG INVARIANCE H_{04dd} MODEL	0.971	0.927337598	0.099112471
PARTIAL STRONG INVARIANCE H_{04ee} MODEL	0.971	0.927381327	0.099261252
DIFF [H_{04a}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.002**	**-0.004983366**	**-0.016168396**
DIFF [H_{04b}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.002**	**-0.004665339**	**-0.015212494**
DIFF [H_{04c}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.002**	**-0.003724435**	**-0.012326024**
DIFF [H_{04d}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.00342004**	**-0.011373243**
DIFF [H_{04e}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.00339104**	**-0.01128198**
DIFF [H_{04f}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.003216999**	**-0.010732475**
DIFF [H_{04g}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.002607341**	**-0.008783038**
DIFF [H_{04h}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.002578289**	**-0.008689183**
DIFF [H_{04i}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.002578289**	**-0.008689183**
DIFF [H_{04j}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.002563763**	**-0.008642220**
DIFF [H_{04k}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.002534709**	**-0.008548223**
DIFF [H_{04l}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.00227314**	**-0.007698014**
DIFF [H_{04m}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.002055052**	**-0.006983637**
DIFF [H_{04n}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**0.001**	**-0.001996878**	**-0.006792230**
DIFF [H_{04o}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.001982334**	**-0.006744318**
DIFF [H_{04p}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.001924151**	**-0.006552432**
DIFF [H_{04q}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.001880509**	**-0.006408265**
DIFF [H_{04r}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.001822313**	**-0.006215706**
DIFF [H_{04s}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.001735005**	**-0.005926144**
DIFF [H_{04t}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.001647681**	**-0.005635713**
DIFF [H_{04u}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.00161857**	**-0.005538709**
DIFF [H_{04v}-H_a;TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.001502105**	**-0.005149720**
DIFF [H_{04w}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.001429299**	**-0.004905811**
DIFF [H_{04x}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**-0.001**	**-0.001443861**	**-0.004954641**
DIFF [H_{04y}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**0**	**-0.001341918**	**-0.004612313**
DIFF [H_{04z}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**0**	**-0.001225383**	**-0.004219612**
DIFF [H_{04aa}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**0**	**-0.001123391**	**-0.003874707**
DIFF [H_{04bb}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**0**	**-0.001108819**	**-0.003825336**
DIFF [H_{04cc}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**0**	**-0.001006802**	**-0.003479049**
DIFF [H_{04dd}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**0**	**-0.001006802**	**-0.0034279049**
DIFF [H_{04ee}-H_a; TEST OF PARTIAL SCALAR EQUIVALENCE]	**0**	**-0.000963073**	**-0.003330269**
MODEL			

CFI= Comparative Fit Index; Γ_1= Gamma Hat fit index; Mc= McDonald Non-centrality index. H_{04a}-H_a = Item WH40 allowed to vary; H_{04b}-H_a = Items WH40 and GEN2 allowed to vary; H_{04c}-H_a = Items WH40, GEN2 and WL47 allowed to vary, H_{04d}-H_a = Items WH40, GEN2, WL47 and GEN3 allowed to vary; H_{04e}-H_a = Items WH40, GEN2, WL47, GEN3 and WH39 allowed to vary; H_{04f}-H_a = Items WH40, GEN2, WL47, GEN3, WH39 and WH42 allowed to vary; H_{04g}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42 and WL48 allowed to vary; H_{04h}-H_a = Items WH40, GEN2, WL47 GEN3, WH39, WH42, WL48, and CA22

allowed to vary; H_{04i}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, and WH41 allowed to vary; H_{04j}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, and WH44 allowed to vary; H_{04k}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, and GEN8 allowed to vary; H_{04l}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, and TE20 allowed to vary; H_{04m}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, and RA1 allowed to vary; H_{04n}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20,RA1, and TE17 allowed to vary; H_{04o}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, and WH37 allowed to vary; H_{04p}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, and GEN1 allowed to vary; H_{04q}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, and WL45 allowed to vary; H_{04r}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WH45, and CA21 allowed to vary; H_{04s}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, and CA23 allowed to vary; H_{04t}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, and RA2 allowed to vary; H_{04u}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2 and WH38 allowed to vary; H_{04v}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38 and LA36 allowed to vary; H_{04w}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36 and RA6 allowed to vary; H_{04x}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36 and GEN4 allowed to vary; H_{04y}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4 and LA32 allowed to vary; H_{04z}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, and REL11 allowed to vary; H_{04aa}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11 and REL13 allowed to vary; H_{04bb}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11, REL13 and WL49 allowed to vary; H_{04cc}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11, REL13, WL49 and REL15 allowed to vary; H_{04dd}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11, REL13, WL49, REL15 and GEN9 allowed to vary; H_{04ee}-H_a = Items WH40, GEN2, WL47, GEN3, WH39, WH42, WL48, CA22, WH41, WH44, GEN8, TE20, RA1, TE17, WH37, GEN1, WL45, CA21, CA23, RA2, WH38, LA36, RA6, GEN4, LA32, REL11, REL13, WL49, REL15, GEN9 and RA5 allowed to vary.

Partial scalar equivalence may be concluded if the partial strong invariance multi-group measurement model in which the three non-invariant factor loadings (LA35, JS27 and RA6) were allowed to vary, and the identified non-invariant intercepts were allowed to vary, does not fit practically significantly poorer than the configural invariance multi-group measurement model. As indicated in Table 5.63, for the partial strong invariance models in which the intercepts of the thirty-one indicator variables on which the intercept estimates differed most across the two gender samples in the configural model (WH40 to RA5 as listed in Table 5.61) were not constrained to be equal across gender groups, the change in CFI, Γ_1 and Mc was less than the critical thresholds of -0.01, -0.001 and -0.02 respectively. This provided support for partial scalar equivalence.

5.15.1 DECISION ON THE RESULTS OF PARTIAL SCALAR EQUIVALENCE

Partial scalar equivalence was adequately supported as a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups and in which all parameters were estimated freely across the gender samples, but for the slopes of the regression of 56 indicator variables on the latent variables and 28 intercepts, did not fit practically significantly poorer than a multi-group measurement model in which the structure of the model was constrained

to be the same across gender groups, but all parameters were estimated freely. When the partial strong invariance model, which allowed for the non-invariant factor loadings of LA35, JS27 and RA6 to vary across gender groups and the 31 non-invariant intercepts to vary across gender groups, was compared to the configural invariance model, partial scalar equivalence was displayed as all three practical fit indices were below the critical thresholds.

Therefore, along with 56 of the 59 SWSI slopes being invariant, 28 of the 59 SWSI intercepts were invariant, thereby displaying adequate support for partial scalar equivalence. This finding implied a partial lack of uniform item bias, when freeing the non-invariant items.

Under these conditions, strict invariance was tested. The multi-group strict invariance model was specified to allow for the 3 non-invariant factor loadings and the 31 non-invariant intercepts to vary across gender groups.

5.16 STRICT INVARIANCE

Upon finding acceptable model fit for the strong invariance model, as well as establishing partial scalar equivalence, strict invariance was tested. This test aimed to establish whether a multi-group measurement model in which the structure of the model was constrained to be same across gender groups, and in which all parameters were estimated freely across the samples, but for the 56 invariant factor loadings, and the 28 invariant regression intercepts, and all the measurement error variances were constrained to be equal across gender groups, demonstrates acceptable fit when fitted to the samples simultaneously in a multi-group analysis.

Operational hypothesis 5 was tested by testing H_{05}: RMSEA ≤ .05[32].

The multi-group SWSI model, in which the structure of the model, the factor loadings (except for the factor loadings of items LA35, JS27 and RA6, which are allowed to vary), the vector of the regression intercepts (except for the 31 non-invariant intercepts), and the measurement error variances of the indicator variables were constrained to be same across gender groups, was fitted to the male (*N*=460) and female (*N*=460) samples. A visual representation of the fitted multi-group strict invariance measurement model is shown in Figure 5.10.

[32] Strictly speaking operational hypothesis5 and H_{05} and the hypotheses tested here are not exactly the same.

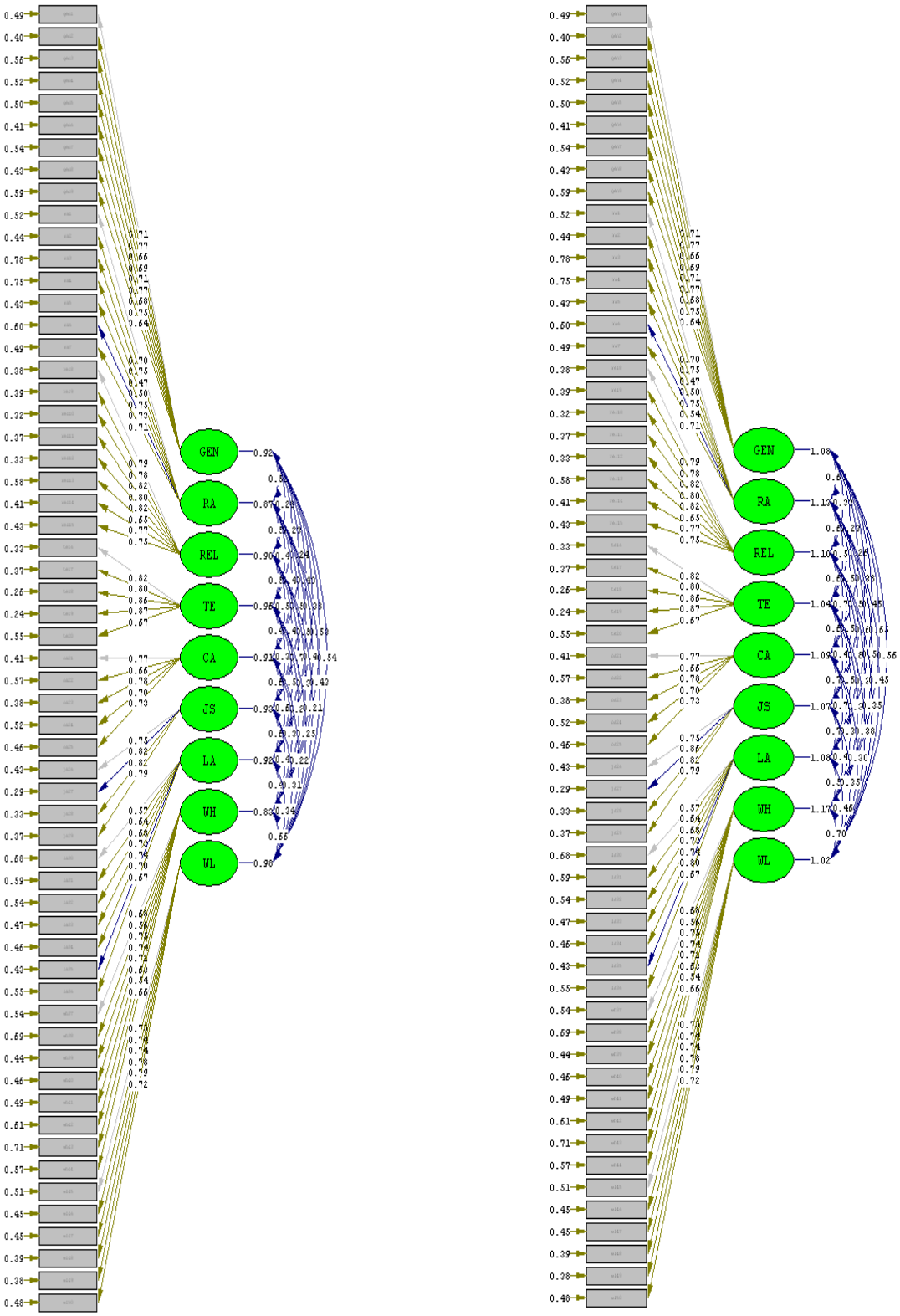

Figure 5.10. Representation of the fitted multi-group SWSI strict invariance measurement model for the male and female samples respectively

5.16.1 MEASUREMENT MODEL FIT INDICES

The strict invariance measurement model converged in 36 iterations. The spectrum of fit statistics is shown in Table 5.64.

Table 5.64

Goodness of fit statistics for the SWSI strict invariance measurement model

Degrees of Freedom = 3366
Minimum Fit Function Chi-Square = 8505.262 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 9191.727 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 7640.720 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 4274.720
90 Percent Confidence Interval for NCP = (4024.460 ; 4532.501)

Minimum Fit Function Value = 9.265
Population Discrepancy Function Value (F0) = 4.657
90 Percent Confidence Interval for F0 = (4.384 ; 4.937)
Root Mean Square Error of Approximation (RMSEA) = 0.0526
90 Percent Confidence Interval for RMSEA = (0.0510 ; 0.0542)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000

Expected Cross-Validation Index (ECVI) = 8.959
90 Percent Confidence Interval for ECVI = (8.558 ; 9.112)
ECVI for Saturated Model = 3.856
ECVI for Independence Model = 161.365

Chi-Square for Independence Model with 3422 Degrees of Freedom = 148015.476
Independence AIC = 148251.476
Model AIC = 8224.720
Saturated AIC = 7080.000
Independence CAIC = 148938.752
Model CAIC = 9925.437
Saturated CAIC = 27698.283

Normed Fit Index (NFI) = 0.948
Non-Normed Fit Index (NNFI) = 0.970
Parsimony Normed Fit Index (PNFI) = 0.933
Comparative Fit Index (CFI) = 0.970
Incremental Fit Index (IFI) = 0.970
Relative Fit Index (RFI) = 0.948

Critical N (CN) = 428.696

Upon first inspection of Table 5.64, the degrees of freedom were indicated to be 3366. This did not correspond with the calculations in Table 4.1, as the non-invariant freed factor loadings and intercepts were not taken into account in the original calculation. A recalculation resulted in the degrees of freedom for this multi-group measurement invariance model, which takes the three non-

invariant factor loadings and the 31 non-invariant intercepts into account, to be 3658-292=3366. This confirms that the model was specified correctly.

The following close fit null hypothesis, which explicitly assumes that the measurement model only approximates the processes that operated in reality to create the observed covariance matrix (Browne & Cudeck, 1993), was tested:

$$H_{05}: \text{RMSEA} \leq .05$$

$$H_{a5}: \text{RMSEA} > .05$$

A RMSEA value of .0526 (and sufficiently small confidence interval: .0510; .0542) was obtained (Table 5.64), indicating that the measurement model showed reasonable to good fit in the sample. The p-value for Test of Close Fit (H_{05}: RMSEA < .05) was 1.00, therefore the close fit null hypothesis H_{05}: RMSEA ≤ .05 was not rejected (p > .05). The position that the measurement model showed close fit was therefore permissible. The measurement model was therefore a plausible explanation for the observed covariance matrices in that the multi-group strict invariance model approximately reproduced the observed covariance matrices, but not perfectly.

5.16.2 DECISION ON THE SUCCESS OF STRICT INVARIANCE

Upon fitting the strict invariance measurement model, it was established that the multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, and in which all parameters were estimated freely across the gender samples, but for the 56 invariant factor loadings, the 28 invariant regression intercepts and all the measurement error variances of the indicator variables that were constrained to be equal, demonstrated close fit when fitted to the gender samples simultaneously in a multi-group analysis.

Strict invariance was thus displayed, which indicated that the position that the 56 regression slopes, the 28 intercepts and all the error variances of the indicator variables were the same across gender samples was plausible. This finding implies that the error variances of the indicator variables are the same across gender samples, suggesting that the respondents from the different gender groups responded to the SWSI in such a manner that no significant variance existed across samples in terms of error terms associated with the indicator variable. The position that a partial lack of item bias has been established was therefore a tenable position. The position that the error variances differ across gender samples could not as yet be claimed to be a less tenable position. To examine this possibility conditional probability equivalence was tested.

5.17 CONDITIONAL PROBABILITY EQUIVALENCE

A finding of strict invariance allowed for the testing of conditional probability equivalence. The test of conditional probability equivalence via multi-group confirmatory factor analysis using LISREL determines whether the multi-group (strict invariance) measurement model in which the 56 elements of $\boldsymbol{\Lambda}^{x}$, the 28 elements of $\boldsymbol{\tau}$ and all the elements of $\boldsymbol{\theta}_{\boldsymbol{\delta}}$ are constrained to be equal across groups fit (a) statistically significantly better or (b) practically significantly better than a multi-group (configural invariance) measurement model in which all model parameters are estimated freely. Whether the multi-group SWSI strict invariance measurement model shows equal probability equivalence is decided by the question whether the strict invariance measurement model fits practically significantly poorer than the configural invariance model. The results of the test whether the multi-group SWSI strict invariance measurement model fits statistically significantly poorer than the configural invariance model are nonetheless also provided.

The results of the test of the statistical significance of the difference in multi-group measurement model fit are shown in Table 5.65.

Table 5.65
Statistical significance of the scaled chi-squared difference statistic: a test of conditional probability equivalence

HYPOTHESIS	SATORRA-BENTLER CHI SQUARE	NORMAL THEORY CHI-SQUARE	DF	Cd	SCALED DIFFERENCE IN S-B CHI-SQUARE	PROB S-B CHI-SQUARE DIFF	PROB SCALED S-B CHI-SQUARE DIFF
Ha CONFIGURAL INVARIANCE MODEL	7412.868	8900.462	3232				
STRICT INVARIANCE H05 MODEL	7640.72	9191.727	3366				
DIFF(H05-Ha) COND PROB EQUIVALENCE	**227.852**		**134**	**1.2558829**	**231.3777813**	**7.73678E-07**	**3.5876E-07**

If the probability of observing the scaled chi-squared difference in a multi-group sample under the null hypothesis of no difference in fit between the strict and configural invariance models in the parameter (H_0: $\chi^2 = 0$) is smaller than or equal to .05, the null hypothesis is rejected in favour of the hypothesis that the fit of the strict and configural invariance models differ in their parameters. As is evident from Table 5.65, the null hypothesis of no difference in model fit was rejected ($p \leq .05$),

which implies a lack of equivalence of measurement error variances across the two gender groups (i.e. lack of conditional probability equivalence) when using statistical significance as the yardstick of equivalence.

The results for the test of practical significance of the difference in multi-group measurement model fit are shown in Table 5.66.

Table 5.66
Practical significance of the CFI, Gamma Hat and MacDonald difference statistic: a test of conditional probability equivalence

MODEL	CFI	Γ_1	Mc
CONFIGURAL INVARIANCE Ha MODEL	0.971	0.928344400	0.102591520
STRICT INVARIANCE H05 MODEL	0.970	0.926842295	0.0974418
DIFFERENCE [H05-Ha; TEST OF COND PROB EQUIVALENCE]	**-0.001**	**-0.001502105**	**-0.005**

CFI= Comparative Fit Index; Γ_1= Gamma Hat fit index; Mc= McDonald Non-centrality index

Conditional probability equivalence may be concluded if the strict invariance multi-group measurement model does not fit practically significantly poorer than the configural invariance multi-group measurement model. As indicated in Table 5.66, the change in CFI and Mc was less than the critical thresholds of -0.01 and -0.02 respectively, therefore the difference in fit was considered practically insignificant. The Gamma Hat fit index was, however, marginally greater than the critical threshold of -0.001, and therefore the difference in fit was considered practically significant. In terms of the decision-rule specified in Chapter 4, conditional probability equivalence could not be concluded as all three criteria were not met. Therefore, it was concluded that a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, and in which all parameters were estimated freely across the gender samples but for 56 factor loadings, 28 intercepts and all error variances that were constrained to be equal across gender groups, fitted practically poorer than a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, but all parameters were estimated freely.

5.17.1 DECISION ON THE RESULTS OF STRICT INVARIANCE AND CONDITIONAL PROBABILITY EQUIVALENCE

Upon fitting the strict invariance multi-group measurement model, the results indicated acceptable model fit, permitting the conclusion of a partial lack of item bias. When comparing the strict

invariance multi-group measurement model to the configural invariance multi-group measurement model (i.e., one with fewer constraints), conditional probability equivalence was not supported, as the difference in the Gamma Hat fit index was greater than the critical threshold of -0.001. Therefore, although the strict invariance multi-group measurement model, in which the structure of the model was constrained to be the same across gender groups and in which all parameters were estimated freely across the gender samples, but for the 56 factor loadings, the 28 intercepts and all the error variances, fitted the data closely, the configural invariance model fitted the data practically significantly better.

Inadequate support for conditional probability equivalence required additional tests to determine the source of non-invariance in terms of measurement error. As conditional probability equivalence, or at least partial conditional probability equivalence, is a necessary prerequisite to conclude partial lack of item bias under a strict interpretation of item bias, accurate identification of non-invariant error variances is necessary.

5.18 PARTIAL CONDITIONAL PROBABILITY EQUIVALENCE

A lack of support for conditional probability equivalence resulted in the need for additional tests to determine the source of non-equivalence, therefore testing for partial conditional probability equivalence. As the three non-invariant slopes (LA35, JS27, and RA6) and 31 non-invariant intercepts (WH40-RA5, refer to Table 5.61) have already been identified, the non-invariant error variances should be correctly identified (Cheung & Rensvold, 1999).

The procedure that was followed involved (a) examining the theta-delta's of the configural invariance model to identify the greatest difference between gender groups (i.e., the most dissimilar error variances), and (b) testing for partial conditional probability equivalence when lifting the error variance equality constraint one item at a time, starting with the item with the most dissimilar error variances across the two gender samples.

All measurement error variances were considered for possible non-invariance. The difference between the male and female error variances in the Theta Delta common metric completely standardised solution was calculated, and the absolute values rank ordered from high to low in Table 5.67 in order to determine the items with the most dissimilar error variances.

Table 5.67

Theta_delta_differences

	Configural invariance: LISREL Estimate Theta_delta		Theta_delta_diff	Rank order from largest to smallest	
	Male	**Female**	**Difference**	**Item**	**Difference**
GEN1	0.510	0.467	0.043	**WH43**	**0.275**
GEN2	0.371	0.437	-0.066	**LA35**	**0.221**
GEN3	0.571	0.539	0.032	**JS27**	**0.165**
GEN4	0.529	0.507	0.022	**REL9**	**0.156**
GEN5	0.426	0.570	-0.144	CA21	0.155
GEN6	0.389	0.417	-0.028	GEN5	0.144
GEN7	0.591	0.485	0.106	RA1	0.139
GEN8	0.393	0.475	-0.082	RA6	0.136
GEN9	0.624	0.561	0.063	CA22	0.119
RA1	0.569	0.430	0.139	WH41	0.115
RA2	0.485	0.374	0.111	RA2	0.111
RA3	0.764	0.805	-0.041	RA4	0.110
RA4	0.696	0.806	-0.110	LA31	0.110
RA5	0.420	0.449	-0.029	GEN7	0.106
RA6	0.532	0.668	-0.136	LA30	0.101
RA7	0.452	0.547	-0.095	RA7	0.095
REL8	0.353	0.410	-0.057	REL12	0.093
REL9	0.313	0.469	-0.156	WH42	0.090
REL10	0.300	0.347	-0.047	WL45	0.088
REL11	0.344	0.392	-0.048	JS29	0.087
REL12	0.370	0.277	0.093	GEN8	0.082
REL13	0.553	0.605	-0.052	WH38	0.081
REL14	0.387	0.428	-0.041	WL49	0.077
REL15	0.397	0.460	-0.063	TE20	0.070
TE16	0.331	0.324	0.007	GEN2	0.066
TE17	0.400	0.337	0.063	REL15	0.063
TE18	0.282	0.237	0.045	TE17	0.063
TE19	0.230	0.252	-0.022	WL50	0.063
TE20	0.518	0.588	-0.070	GEN9	0.063
CA21	0.485	0.330	0.155	JS26	0.061
CA22	0.631	0.512	0.119	LA32	0.059
CA23	0.366	0.393	-0.027	REL8	0.057
CA24	0.540	0.493	0.047	LA34	0.057
CA25	0.460	0.461	-0.001	REL13	0.052
JS26	0.392	0.453	-0.061	LA36	0.052
JS27	0.369	0.204	0.165	REL11	0.048
JS28	0.346	0.313	0.033	CA24	0.047
JS29	0.328	0.415	-0.087	REL10	0.047

LA30	0.723	0.622	0.101	TE18	0.045
LA31	0.532	0.642	-0.110	GEN1	0.043
LA32	0.564	0.505	0.059	RA3	0.041
LA33	0.459	0.471	-0.012	REL14	0.041
LA34	0.483	0.426	0.057	WH40	0.038
LA35	0.320	0.541	-0.221	WL47	0.034
LA36	0.527	0.579	-0.052	JS28	0.033
WH37	0.539	0.532	0.007	GEN3	0.032
WH38	0.646	0.727	-0.081	WL46	0.031
WH39	0.430	0.455	-0.025	RA5	0.029
WH40	0.437	0.475	-0.038	GEN6	0.028
WH41	0.432	0.547	-0.115	CA23	0.027
WH42	0.562	0.652	-0.090	WH39	0.025
WH43	0.570	0.845	-0.275	GEN4	0.022
WH44	0.576	0.561	0.015	TE19	0.022
WL45	0.465	0.553	-0.088	WL48	0.016
WL46	0.458	0.427	0.031	WH44	0.015
WL47	0.434	0.468	-0.034	LA33	0.012
WL48	0.384	0.400	-0.016	TE16	0.007
WL49	0.343	0.420	-0.077	WH37	0.007
WL50	0.449	0.512	-0.063	CA25	0.001

GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface. **Items with non-invariant error variances are in boldface**.

The process to identify non-invariant error variances continued with the performance of a series of tests. Using the rank ordered absolute values, in which the items with the greatest difference in $\boldsymbol{\theta_\delta}$ between gender groups had been identified, a multi-group model was estimated in which all error variances were constrained to be equal, but for the item that obtained the most dissimilar error variances in the fully unconstrained solution, while still honouring the results of the tests for partial scalar equivalence. Under a further finding of a lack of partial conditional probability equivalence the process was repeated by subsequently freeing the item that had the second-most dissimilar error variance in the fully unconstrained solution. The process of freeing non-invariant error variances (adding a non-invariant error variance each time) continued until practical insignificance was achieved, indicating that a partial strict invariance multi-group measurement model has been found that acknowledged the items with the non-invariant error variances, and that did not fit practically significantly poorer than the configural invariance multi-group measurement model.

Following this procedure, the statistical significance of the scaled Satorra-Bentler chi-square difference statistic (Table 5.68) was calculated for each of the partial strict invariance multi-group models in which the most dissimilar measurement error variances were allowed to vary across gender samples until statistical significance was established.

Upon fitting each multi-group partial strict invariance model per non-invariant stress item, the degrees of freedom were examined. The degrees of freedom tally with the calculations[33] confirming that the models were estimated correctly.

Table 5.68

Statistical significance of the scaled chi-squared difference statistic: a test of partial conditional probability equivalence per error variance

HYPOTHESIS	SATORRA-BENTLER CHI SQUARE	NORMAL THEORY CHI-SQUARE	DF	Cd	SCALED DIFFERENCE IN S-B CHI-SQUARE	PROB S-B CHI-SQUARE DIFF	PROB SCALED S-B CHI-SQUARE DIFF
H_a CONFIGURAL INVARIANCE MODEL	7412.868	8900.462	3232				
PARTIAL STRICT INVARIANCE H_{05a} MODEL	7628.536	9172.376	3365				
PARTIAL STRICT INVARIANCE H_{05b} MODEL	7617.214	9155.367	3364				
PARTIAL STRICT INVARIANCE H_{05c} MODEL	7609.466	9140.301	3363				
PARTIAL STRICT INVARIANCE H_{05d} MODEL	7603.362	9126.404	3362				
DIFF(H_{05a}-H_a) PARTIAL COND PROB EQUIVALENCE	**215.668**		**133**	**1.243681**	**218.6363939**	**7.59423E-06**	**4.1361E-06**
DIFF(H_{05b}-H_a) PARTIAL COND PROB EQUIVALENCE	**204.346**		**132**	**1.232633**	**206.7971046**	**5.47958E-05**	**3.4291E-05**
DIFF(H_{05c}-H_a) PARTIAL COND PROB EQUIVALENCE	**196.598**		**131**	**1.213458**	**197.6492733**	**0.000183009**	**0.00015127**
DIFF(H_{05d}-H_a) PARTIAL COND PROB EQUIVALENCE	**190.494**		**130**	**1.191222**	**189.6724229**	**0.000434617**	**0.00050083**

H_{05a}-H_a = Item WH43 allowed to vary; H_{05b}-H_a = Items WH43 and LA35 allowed to vary; H_{05c}-H_a = Items WH43, LA35 and JS27 allowed to vary; H_{05d}-H_a = Items WH43, LA35, JS27 and REL9 allowed to vary.

[33] Degrees of freedom calculations for each partial strict invariance model were calculated separately by the researcher in order to ensure each model was specified correctly.

If the probability of observing the scaled chi-squared difference in a multi-group sample under the null hypothesis of no difference in fit between the partial strict invariance model and the configural invariance model in the parameter (H_0: $\chi^2 = 0$) is smaller than or equal to .05, the null hypothesis of no difference in fit in the parameter is rejected in favour of the hypothesis that the partial strict invariance and configural invariance models do differ in the parameter. As can be seen in Table 5.68, the null hypothesis of no difference in the parameters was rejected ($p \leq .05$) for each of the partial strict invariance models that were compared to the configural invariance model. This implies a lack of equivalence of those error variances constrained to be equal across the two gender groups even when the four most dissimilar error variances are allowed to vary (i.e., a lack of partial conditional probability equivalence) when using statistical significance as the yardstick of equivalence.

The results for the test of practical significance of the difference in multi-group measurement model fit, are shown in Table 5.69.

Table 5.69

Practical significance of the CFI, Gamma Hat and MacDonald difference statistic: a test of partial conditional probability equivalence per error variance

MODEL	CFI	Γ_1	Mc
CONFIGURAL INVARIANCE Ha MODEL	0.971	0.9283444	0.10259152
PARTIAL STRICT INVARIANCE H05a MODEL	0.971	0.927031613	0.098077235
PARTIAL STRICT INVARIANCE H05b MODEL	0.971	0.927191866	0.098618146
PARTIAL STRICT INVARIANCE H05c MODEL	0.971	0.927293874	0.098963914
PARTIAL STRICT INVARIANCE H05d MODEL	0.971	0.927381327	0.099261252
DIFF [H05a-Ha; TEST OF PARTIAL COND PROB EQUIVALENCE]	**0**	**-0.001312787**	**-0.004514285**
DIFF [H05b-Ha; TEST OF PARTIAL COND PROB EQUIVALENCE]	**0**	**-0.001152534**	**-0.003973374**
DIFF [H05c-Ha; TEST OF PARTIAL COND PROB EQUIVALENCE]	**0**	**-0.001050526**	**-0.003627606**
DIFF [H05d-Ha; TEST OF PARTIAL COND PROB EQUIVALENCE]	**0**	**-0.000963073**	**-0.003330269**

CFI= Comparative Fit Index; Γ_1= Gamma Hat fit index; Mc= McDonald Non-centrality index. H_{05a}-H_a = Item WH43 allowed to vary; H_{05b}-H_a = Items WH43 and LA35 allowed to vary; H_{05c}-H_a = Items WH43, LA35 and JS27 allowed to vary; H_{05d}-H_a = Items WH43, LA35, JS27 and REL9 allowed to vary.

Partial conditional probability equivalence could be concluded if the partial strict invariance multi-group measurement model in which the four non-invariant factor loadings (LA35, JS27 and RA6) , the 31 non-invariant intercepts, and the non-invariant error variances were allowed to vary, did not fit practically significantly poorer than the configural invariance multi-group measurement model. As indicated in Table 5.69, for the partial conditional probability invariance models in which the error variances of the indicator variables WH43, LA35, JS27 and REL9 were not constrained to be equal across gender groups, the change in CFI, Γ_1 and Mc was less than the critical thresholds of -0.01, -

0.001 and -0.02 respectively. This provides support for partial conditional probability equivalence when using practical significance as the yardstick of equivalence.

5.18.1 DECISION ON THE RESULTS OF PARTIAL CONDITIONAL PROBABILITY EQUIVALENCE

Partial conditional probability equivalence would adequately be supported when a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups and in which all parameters were estimated freely across the gender samples, but for the 56 slopes, 28 intercepts, and 55 error variances of the indicator variables which were constrained to be equal, did not fit practically significantly poorer than a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, but all parameters were estimated freely. When the partial strict invariance model, which allowed for the non-invariant slopes of LA35, JS27 and RA6, the 31 non-invariant intercepts, and the non-invariant error variances of WH43, LA35, JS27 and REL9 to vary across gender groups, was compared to the configural invariance model, partial conditional probability equivalence was displayed as all three practical fit indices were below the critical thresholds.

Therefore, along with 56 of the 59 SWSI slopes and 28 of the 59 SWSI intercepts being invariant, 55 of the 59 error variances of the indicator variables were invariant, displaying adequate support for partial conditional probability equivalence. This means that a sufficient number of items variables have equal error variances across groups, which suggests that the scale reliabilities can be compared across groups (Cheung & Rensvold, 1999). This finding implied a lack of item bias for those invariant items.

Under these conditions, complete invariance was tested.

5.19 COMPLETE INVARIANCE

Upon finding acceptable model fit for the strict invariance model, as well as partial conditional probability equivalence, complete invariance was tested. This test aimed to whether a multi-group measurement model in which the structure of the model was constrained to be same across gender groups and in which all parameters were estimated freely across the samples, but for the 56 invariant factor loadings, the 28 invariant regression intercepts and the 55 measurement error variances, and all latent variable variances and covariances were constrained to be equal across gender groups, demonstrated acceptable fit when fitted to the samples simultaneously in a multi-group analysis.

Operational hypothesis 6 was tested by testing H_{06}: RMSEA ≤ .05[34].

The SWSI model, in which the structure of the model, the factor loadings (except for the factor loadings of items LA35, JS27 and RA6, which are allowed to vary), the vector of the regression intercepts (except for the 31 non-invariant intercepts), and the measurement error variances of the indicator variables were constrained to be same across gender groups (except for the error variances of items WH43, LA35, JS27, and REL9, which are allowed to vary), and in which all the latent variable variances and covariances were constrained to be the same across gender groups was fitted to the male (N=460) and female (N=460) samples. A visual representation of the fitted multi-group complete invariance model is shown in Figure 5.11.

[34] Strictly speaking operational hypothesis 6 and H_{06} and the hypotheses tested here are not exactly the same.

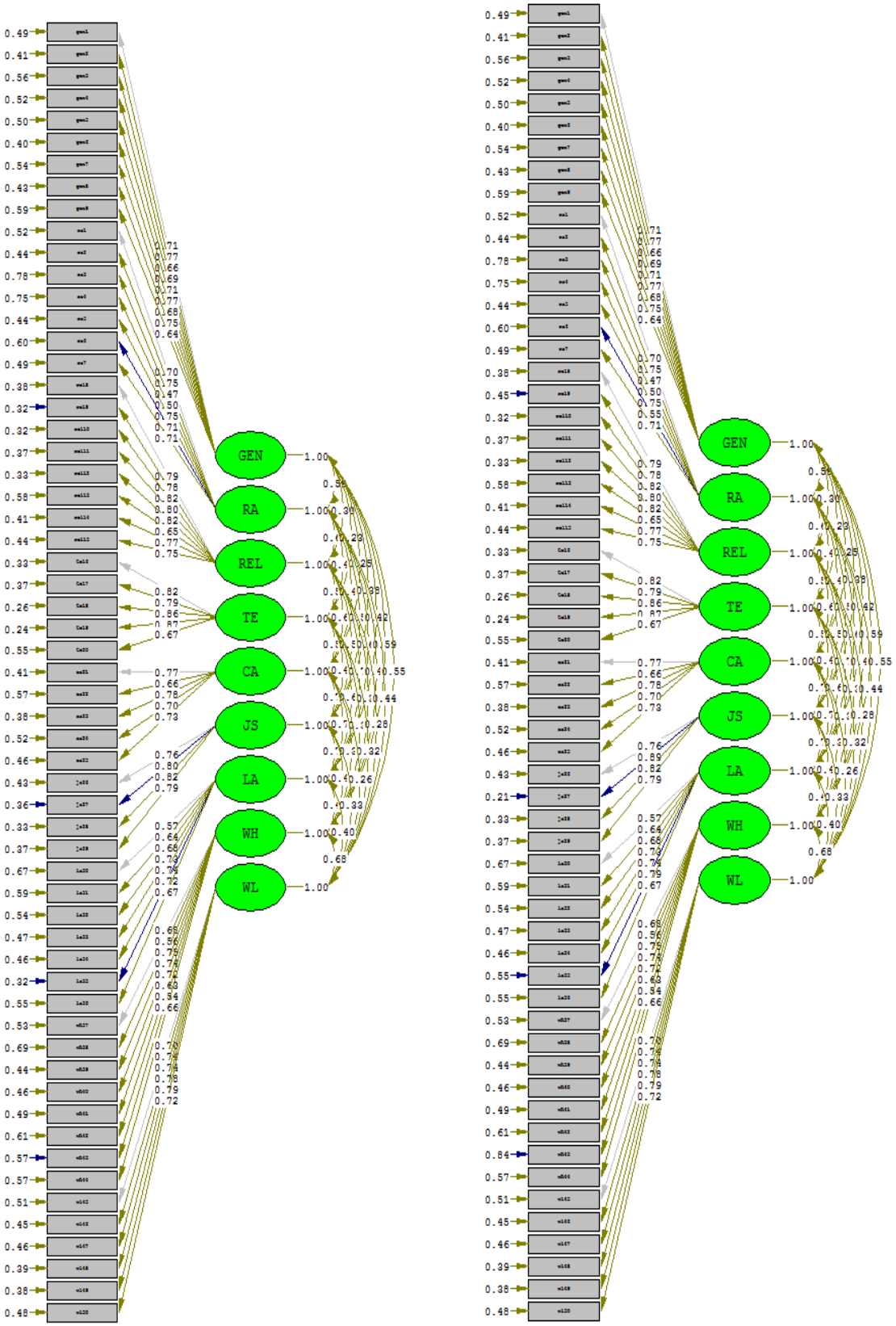

Figure 5.11. Representation of the fitted multi-group SWSI complete invariance measurement model for the male and female samples respectively

5.19.1 MEASUREMENT MODEL FIT INDICES

The complete invariance measurement model converged in 34 iterations. The spectrum of fit statistics is shown in Table 5.70

Table 5.70

Goodness of fit statistics for the SWSI complete invariance measurement model

Degrees of Freedom = 3407
Minimum Fit Function Chi-Square = 8506.391 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 9220.150 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 7679.727 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 4272.727
90 Percent Confidence Interval for NCP = (4022.075 ; 4530.905)

Minimum Fit Function Value = 9.266
Population Discrepancy Function Value (F0) = 4.654
90 Percent Confidence Interval for F0 = (4.381 ; 4.936)
Root Mean Square Error of Approximation (RMSEA) = 0.0523
90 Percent Confidence Interval for RMSEA = (0.0507 ; 0.0538)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000

Expected Cross-Validation Index (ECVI) = 8.913
90 Percent Confidence Interval for ECVI = (8.511 ; 9.065)
ECVI for Saturated Model = 3.856
ECVI for Independence Model = 161.365

Chi-Square for Independence Model with 3422 Degrees of Freedom = 148015.476
Independence AIC = 148251.476
Model AIC = 8181.727
Saturated AIC = 7080.000
Independence CAIC = 148938.752
Model CAIC = 9643.645
Saturated CAIC = 27698.283

Normed Fit Index (NFI) = 0.948
Non-Normed Fit Index (NNFI) = 0.970
Parsimony Normed Fit Index (PNFI) = 0.944
Comparative Fit Index (CFI) = 0.970
Incremental Fit Index (IFI) = 0.970
Relative Fit Index (RFI) = 0.948

Critical N (CN) = 431.563

The degrees of freedom for this model were 3407 (see Table 5.70). This did not correspond with the calculations in Table 4.1, as the non-invariant freed factor loadings, intercepts and error variances were not taken into account in the original calculation. A recalculation resulted in the degrees of freedom for this multi-group measurement invariance model, which took the three non-invariant

factor loadings, the 31 non-invariant intercepts, and the four non-invariant error terms into account, to be 3658-251=3407. This confirmed that the model was specified correctly.

The following close fit null hypothesis, which explicitly assumes that the measurement model only approximates the processes that operated in reality to create the observed covariance matrix (Browne & Cudeck, 1993), was tested:

$$H_{06}: \text{RMSEA} \leq .05$$

$$H_{a6}: \text{RMSEA} > .05$$

A RMSEA value of .0523 was obtained (small confidence interval of .0507; .0538), indicating that the measurement model showed reasonable to good fit. The p-value for Test of Close Fit (H_{05}: RMSEA < .05) was 1.00, therefore the close fit null hypothesis H_{05}: RMSEA ≤ .05 was not rejected (p > .05). The position that the measurement model showed close fit was therefore permissible. The measurement model therefore provided a plausible explanation for the observed covariance matrices in that the model approximately reproduced the observed covariance matrices, but not perfectly.

5.19.2 DECISION ON THE SUCCESS OF COMPLETE INVARIANCE

Upon fitting the complete invariance measurement model, it was established that the multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, and in which all parameters were constrained across the gender samples, but for the three non-invariant factor loadings, the 31 non-invariant regression intercepts, and the 4 non-invariant error variances of the indicator variables that were allowed to vary, demonstrated acceptable fit when fitted to the gender samples simultaneously in a multi-group analysis.

Complete invariance was displayed, which indicated that the position that the 56 regression slopes, the 28 intercepts, the 55 error variances of the indicator variables, and all the latent variable variances and covariances were the same across gender samples was tenable. This finding implies that the latent variable variances and covariances are the same across gender samples. Support was thus provided for the fact that the participants from the different gender samples used equivalent ranges of the construct continuum to respond to the indicators reflecting the construct (Vandenberg & Lance, 2000).

5.20 FULL EQUIVALENCE

A finding of complete invariance allowed for the testing of full equivalence. The test of full equivalence via multi-group confirmatory factor analysis using LISREL determines whether the multi-

group (complete invariance) measurement model in which the 56 parameters in $\boldsymbol{\Lambda^x}$, the 28 parameters in $\boldsymbol{\tau}$, the 55 parameters in the main-diagonal of $\boldsymbol{\Theta_\delta}$, and all the parameters in $\boldsymbol{\Phi}$ are constrained to be equal across groups fits (a) statistically significantly better or (b) practically significantly better than a multi-group (configural invariance) measurement model in which all model parameters are estimated freely.

The results of the test of the statistical significance of the difference in multi-group measurement model fit, are shown in Table 5.71

Table 5.71

Statistical significance of the scaled chi-squared difference statistic: a test of full equivalence

HYPOTHESIS	SATORRA-BENTLER CHI SQUARE	NORMAL THEORY CHI-SQUARE	DF	Cd	SCALED DIFFERENCE IN S-B CHI-SQUARE	PROB S-B CHI-SQUARE DIFF	PROB SCALED S-B CHI-SQUARE DIFF
Ha CONFIGURAL INVARIANCE MODEL	7412.868	8900.462	3232				
COMPLETE INVARIANCE MODEL H06	7679.727	9220.15	3407				
DIFF(H06-Ha) FULL EQUIVALENCE	**266.859**		**175**	**0.696495**	**458.9950693**	**9.37213E-06**	**2.44346E-27**

If the probability of observing the scaled chi-squared difference in a multi-group sample under the null hypothesis of no difference in fit between the complete and configural invariance models in the parameter (H_0: $\chi = 0$) is smaller than or equal to .05, the null hypothesis is rejected in favour of the hypothesis that the fit of the complete and configural invariance models do differ in the parameter. As can be seen in Table 5.71, the null hypothesis of no difference in multi-group model fit in the parameter was rejected ($p \leq .05$). This implied lack of equivalence of latent variable variances and covariances across the two gender groups (i.e., lack of full equivalence) when using statistical significance as the yardstick of equivalence.

The results for the test of practical significance of the difference in multi-group measurement model fit are shown in Table 5.72.

Table 5.72

Practical significance of the CFI, Gamma Hat and MacDonald difference statistic: a test of full equivalence

MODEL	CFI	Γ_1	Mc
CONFIGURAL INVARIANCE Ha MODEL	0.971	0.9283444	0.10259152
COMPLETE INVARIANCE H06 MODEL	0.971	0.927104449	0.098322735
DIFF [H06-Ha; TEST OF FULL EQUIVALENCE]	**-0.001**	**-0.001458423**	**-0.005**

CFI= Comparative Fit Index; Γ_1= Gamma Hat fit index; Mc= Mcdonald Non-centrality index

Full equivalence may be concluded if the complete invariance multi-group measurement model does not fit practically significantly poorer than the configural invariance multi-group measurement model. As indicated in Table 5.72, the change in CFI and Mc was less than the critical thresholds of -0.01 and -0.02 respectively. The difference in the Gamma Hat fit index was, however, marginally greater than the critical threshold of -0.001, and therefore considered practically significant. Given the decision rule specified in Chapter 4, full equivalence cannot be concluded as all three criteria were not met. Therefore, a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, and in which all parameters were constrained to be the same across the gender samples, but for three non-invariant factor loadings, 31 non-invariant intercepts, and four non-invariant error variances which were allowed to vary, fitted practically significantly poorer than a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups but all parameters were estimated freely.

5.20.1 DECISION ON THE RESULTS OF COMPLETE INVARIANCE AND FULL EQUIVALENCE

Upon fitting the complete invariance multi-group measurement model, the results indicated acceptable model fit, permitting the conclusion of partial measurement invariance across gender samples. When comparing the complete invariance multi-group measurement model to the configural invariance multi-group measurement model (i.e., one with fewer constraints), full equivalence was not supported as the difference in the Gamma Hat fit index was greater than the critical threshold of -0.001. Therefore, although the complete invariance multi-group measurement model, in which the structure of the model was constrained to be the same across gender groups and in which all parameters were constrained across the gender samples, but for the three factor loadings, the 31 intercepts and the four error variances, fitted the data closely, the configural invariance multi-group model could practically significantly better account for the observed

covariance matrices. Inadequate support for full equivalence required additional tests to determine the source of non-invariance in terms of invariant latent variable variances and covariances.

5.21 PARTIAL FULL EQUIVALENCE

A lack of support for full equivalence resulted in the need for additional tests to determine the sources of non-invariance, resulting in a test for partial full equivalence. As the four non-invariant slopes (LA35, JS27, and RA6), 31 non-invariant intercepts (WH40-RA5, refer to Table 5.61) and four non-invariant error variances (WH43, LA35, JS27, and REL9) have already been identified, the non-invariant variances and covariances had to be correctly identified (Cheung & Rensvold, 1999).

The procedure that was followed involved (a) examining the variance and covariances in the phi matrix of the configural invariance model to identify the greatest difference between gender groups, and (b) testing for partial full equivalence when lifting the latent variable variance and/or covariance equality constraint one item at a time, starting with the item with the most dissimilar latent variable variances or covariances across the two gender samples.

All latent variable variances were considered for possible non-invariance. The difference between the male and female variances in the phi matrix common metric completely standardised solution was calculated and the absolute values rank ordered from high to low in Table 5.73 in order to identify the most dissimilar variances.

Table 5.73

Variance differences

	Configural invariance: Variance (Phi matrix)		**Variance_diff**	**Rank order from largest to smallest**	
Scale	**Male**	**Female**	**Difference**	**Scale**	**Difference**
GEN	0.974	1.026	-0.052	**RA**	**0.740**
RA	0.630	1.370	-0.740	**CA**	**0.320**
REL	0.980	1.020	-0.040	**LA**	**0.188**
TE	1.039	0.961	0.078	**WH**	**0.156**
CA	0.840	1.160	-0.320	**JS**	**0.116**
JS	0.942	1.058	-0.116	**TE**	**0.078**
LA	0.906	1.094	-0.188	**GEN**	**0.052**
WH	0.922	1.078	-0.156	**REL**	**0.040**
WL	1.017	0.983	0.034	**WL**	**0.034**

GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface. **Scales with non-invariant variances are in boldface**.

All covariances were considered for possible non-invariance. The difference between the male and female covariances in the phi matrix common metric completely standardised solution was calculated and the absolute values rank ordered from high to low in Table 5.74 in order to identify the most dissimilar covariances.

Table 5.74

Covariance differences

	Configural invariance: Covariance (Phi matrix)		**Covariance_diff**	**Rank order from largest to smallest**	
Scales	**Male**	**Female**	**Difference**	**Item**	**Difference**
GEN-RA	0.465	0.679	-0.214	**RA-LA**	**0.238**
GEN-REL	0.291	0.311	-0.020	**CA-LA**	**0.238**
GEN-TE	0.251	0.216	0.035	**GEN-RA**	**0.214**
GEN-CA	0.235	0.261	-0.026	**RA-CA**	**0.200**
GEN-JS	0.406	0.363	0.043	**CA-JS**	**0.188**
GEN-LA	0.377	0.455	-0.078	**TE-CA**	**0.183**
GEN-WH	0.568	0.611	-0.043	**REL-CA**	**0.181**
GEN-WL	0.567	0.529	0.038	**RA-REL**	**0.172**
RA-REL	0.492	0.664	-0.172	**RA-TE**	**0.152**
RA-TE	0.389	0.541	-0.152	**RA-JS**	**0.135**
RA-CA	0.348	0.548	-0.200	**LA-WH**	**0.123**
RA-JS	0.439	0.574	-0.135	**LA-WL**	**0.115**
RA-LA	0.493	0.731	-0.238	**REL-WL**	**0.114**
RA-WH	0.428	0.506	-0.078	**JS-LA**	**0.111**
RA-WL	0.383	0.466	-0.083	**TE-WL**	**0.095**
REL-TE	0.574	0.594	-0.020	CA-WL	0.092
REL-CA	0.518	0.699	-0.181	RA-WL	0.083
REL-JS	0.490	0.528	-0.038	GEN-LA	0.078
REL-LA	0.733	0.807	-0.074	RA-WH	0.078
REL-WH	0.350	0.323	0.027	REL-LA	0.074
REL-WL	0.221	0.335	-0.114	TE-LA	0.052
TE-CA	0.416	0.599	-0.183	CA-WH	0.049
TE-JS	0.406	0.423	-0.017	GEN-JS	0.043
TE-LA	0.605	0.657	-0.052	GEN-WH	0.043
TE-WH	0.305	0.340	-0.035	REL-JS	0.038
TE-WL	0.265	0.360	-0.095	GEN-WL	0.038
CA-JS	0.614	0.802	-0.188	GEN-TE	0.035
CA-LA	0.583	0.821	-0.238	TE-WH	0.035
CA-WH	0.316	0.365	-0.049	WH-WL	0.028
CA-WL	0.213	0.305	-0.092	REL-WH	0.027
JS-LA	0.641	0.752	-0.111	GEN-CA	0.026
JS-WH	0.431	0.437	-0.006	GEN-REL	0.020
JS-WL	0.322	0.329	-0.007	REL-TE	0.020

LA-WH	0.417	0.540	-0.123	TE-JS	0.017
LA-WL	0.343	0.458	-0.115	JS-WL	0.007
WH-WL	0.695	0.667	0.028	JS-WH	0.006

GEN = General Work Stress; RA = Role Ambiguity; REL = Relationships; TE = Tools and Equipment; CA = Career Advancement; JS = Job Security; LA = Lack of Autonomy; WH = Work/Home Interface. **Scales with non-invariant covariance are in boldface.**

The process to identify non-invariant variances and covariances continued with the performance of a series of tests. Using the rank ordered absolute values, in which the items with the greatest difference in variance and covariance between gender groups was identified, a multi-group model was estimated in which all variances and covariances were constrained to be equal, but for the latent variable with the most dissimilar variance and the latent variable pair with the most dissimilar covariance in the fully unconstrained solution. Under a finding of a lack of partial full equivalence for this model, the process was repeated by subsequently freeing the second-most dissimilar variance and the second-most dissimilar covariance in the fully unconstrained solution. The process of freeing non-invariant variances and covariances (adding a non-invariant variance and covariance each time) continued until practical insignificance was achieved, indicating that a partial complete invariance multi-group measurement model was found, which acknowledged the latent variables and latent variable pairs with non-invariant variances and covariances, and that did not fit practically significantly poorer than the configural invariance multi-group measurement model.

Following this procedure, the statistical significance (Table 5.75) was calculated for each separate partial complete invariance model that freed an additional variance and covariance term until practical significance was established.

Upon fitting each multi-group partial complete invariance model that freed an additional variance and covariance term, the degrees of freedom were examined. The degrees of freedom tallied with the calculations[35] confirming that the models were estimated correctly.

[35] The degrees of freedom calculations for each partial complete invariance model were calculated separately by the researcher in order to ensure each model was specified correctly.

Table 5.75

Statistical significance of the scaled chi-squared difference statistic: a test of partial full equivalence per variance and covariance

HYPOTHESIS	SATORRA-BENTLER CHI SQUARE	NORMAL THEORY CHI-SQUARE	DF	cd	SCALED DIFFERENCE IN S-B CHI-SQUARE	PROB S-B CHI-SQUARE DIFF	PROB SCALED S-B CHI-SQUARE DIFF
H_a CONFIGURAL INVARIANCE MODEL	7412.868	8900.462	3232				
PARTIAL COMPLETE INVARIANCE MODEL H_{06a}	7665.679	9202.943	3404[36]				
PARTIAL COMPLETE INVARIANCE MODEL H_{06}b	7661.844	9198.97	3402				
PARTIAL COMPLETE INVARIANCE MODEL H_{06c}	7660.065	9196.466	3400				
PARTIAL COMPLETE INVARIANCE MODEL H_{06d}	7645.752	9177.112	3398				
PARTIAL COMPLETE INVARIANCE MODEL $H0_{6e}$	7643.614	9174.386	3396				
PARTIAL COMPLETE INVARIANCE MODEL H_{06f}	7642.462	9172.311	3394				
PARTIAL COMPLETE INVARIANCE MODEL H_{06g}	7638.555	9166.613	3392				
PARTIAL COMPLETE INVARIANCE MODEL H_{06h}	7634.77	9162.590	3390				
PARTIAL COMPLETE INVARIANCE MODEL H_{06i}	7634.895	9162.714	3388				
PARTIAL COMPLETE INVARIANCE MODEL H_{06j}	7635.071	9162.956	3387				
PARTIAL COMPLETE INVARIANCE MODEL H_{06k}	7634.925	9162.997	3386				
PARTIAL COMPLETE INVARIANCE MODEL H_{06l}	7630.024	9156.563	3385				

[36] Three degrees of freedom were lost because both the RA-LA and the CA-LA covariances were freed along with the RA variance in the first partial complete invariance model.

PARTIAL COMPLETE INVARIANCE MODEL H_{06m}	7629.678	9155.869	3384				
PARTIAL COMPLETE INVARIANCE MODEL H_{06n}	7622.720	9147.745	3383				
DIFF(H_{06a}-H_a) PARTIAL FULL EQUIV	**252.811**		**172**	**1.1979318**	**252.5026883**	**5.87681E-05**	**6.19454E-05**
DIFF(H_{06b}-H_a) PARTIAL FULL EQUIV	**248.976**		**170**	**1.199549395**	**248.8501109**	**7.48963E-05**	**7.65154E-05**
DIFF(H_{06c}-H_a) PARTIAL FULL EQUIV	**247.197**		**168**	**1.198564121**	**246.9655104**	**6.721E-05**	**6.99304E-05**
DIFF(H_{06d}-H_a) PARTIAL FULL EQUIV	**232.884**		**166**	**1.192729667**	**231.9469429**	**0.000476273**	**0.00055125**
DIFF(H_{06e}-H_a) PARTIAL FULL EQUIV	**230.746**		**164**	**1.192204614**	**229.7625732**	**0.000459652**	**0.000536418**
DIFF(H_{06f}-H_a) PARTIAL FULL EQUIV	**229.594**		**162**	**1.190207257**	**228.4047576**	**0.000378763**	**0.000457797**
DIFF(H_{06g}-H_a) PARTIAL FULL EQUIV	**225.687**		**160**	**1.187282516**	**224.1682131**	**0.0004839**	**0.000614905**
DIFF(H_{06h}-H_a) PARTIAL FULL EQUIV	**221.902**		**158**	**1.188579969**	**220.5388**	**0.00060568**	**0.000749832**
DIFF(H_{06i}-H_a) PARTIAL FULL EQUIV	**222.027**		**156**	**1.188358105**	**220.6843198**	**0.000406467**	**0.000504557**
DIFF(H_{06j}-H_a) PARTIAL FULL EQUIV	**222.203**		**155**	**1.18837038**	**223.8856804**	**0.000325197**	**0.000403288**
DIFF(H_{06k}-H_a) PARTIAL FULL EQUIV	**222.057**		**154**	**1.188916782**	**220.0818651**	**0.000273435**	**0.000335561**
DIFF(H_{06l}-H_a) PARTIAL FULL EQUIV	**217.156**		**153**	**1.187242525**	**215.7107707**	**0.000501936**	**0.000632767**
DIFF(H_{06m}-H_a) PARTIAL FULL EQUIV	**216.81**		**152**	**1.186344675**	**215.2890349**	**0.000437792**	**0.00055994**
DIFF(H_{06n}-H_a) PARTIAL FULL EQUIV	**209.852**		**151**	**1.186917767**	**208.3404654**	**0.001096034**	**0.001382585**

If the probability of observing the scaled chi-squared difference in a multi-group sample under the null hypothesis of no difference in fit between the partial complete and configural invariance models in the parameter (H0: $\chi^2 = 0$) is smaller than or equal to .05, the null hypothesis is rejected in favour of the hypothesis that the fit of the partial complete and configural invariance models do differ in the parameter. As can be seen in Table 5.75, all 14 tested hypotheses were rejected ($p \leq .05$). This implied a lack of equivalence of latent variable variances and covariances across the two gender groups even when the non-invariant variances and covariances were allowed to vary (i.e., lack of partial full equivalence) when using statistical significance as the yardstick of equivalence.

The results for the test of practical significance of the difference in multi-group measurement model fit, is shown in Table 5.76 in order to not solely rely on statistical significance to decide on partial measurement equivalence.

Table 5.76

Practical significance of the CFI, Gamma Hat and MacDonald difference statistic: a test of partial full equivalence per variance and covariance

MODEL	CFI	Γ_1	Mc
CONFIGURAL INVARIANCE H_a MODEL	0.971	0.928344400	0.10259152
PARTIAL COMPLETE INVARIANCE H_{06a}	0.971	0.927060746	0.098175361
PARTIAL COMPLETE INVARIANCE H_{06b}	0.971	0.927089881	0.098273586
PARTIAL COMPLETE INVARIANCE H_{06c}	0.971	0.927075313	0.098224461
PARTIAL COMPLETE INVARIANCE H_{06d}	0.971	0.927279300	0.098914444
PARTIAL COMPLETE INVARIANCE H_{06e}	0.971	0.927279300	0.098914444
PARTIAL COMPLETE INVARIANCE H_{06f}	0.971	0.927264726	0.098865000
PARTIAL COMPLETE INVARIANCE H_{06g}	0.971	0.927293874	0.098963914
PARTIAL COMPLETE INVARIANCE H_{06h}	0.971	0.927323023	0.099062927
PARTIAL COMPLETE INVARIANCE H_{06i}	0.971	0.927293874	0.098963914
PARTIAL COMPLETE INVARIANCE H_{06j}	0.971	0.927264726	0.098865000
PARTIAL COMPLETE INVARIANCE H_{06k}	0.971	0.927264726	0.098865000
PARTIAL COMPLETE INVARIANCE H_{06l}	0.971	0.927323023	0.099062927
PARTIAL COMPLETE INVARIANCE H_{06m}	0.971	0.927308448	0.099013408
PARTIAL COMPLETE INVARIANCE H_{06n}	0.971	0.927410481	0.099360562
DIFF [H_{06a}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001283654**	**-0.004416159**
DIFF [H_{06b}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001254519**	**-0.004317934**
DIFF [H_{06c}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001269087**	**-0.004367059**
DIFF [H_{06d}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001065100**	**-0.003677076**
DIFF [H_{06e}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001065100**	**-0.003677076**
DIFF [H_{06f}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001079674**	**-0.003726520**
DIFF [H_{06g}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001050526**	**-0.003627606**
DIFF [H_{06h}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001021377**	**-0.003528593**
DIFF [H_{06i}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001050526**	**-0.003627606**
DIFF [H_{06j}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001079674**	**-0.003726520**
DIFF [H_{06k}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001079674**	**-0.003726520**
DIFF [H_{06l}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001021377**	**-0.003528593**
DIFF [H_{06m}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.001035952**	**-0.003578112**
DIFF [H_{06n}-H_a; TEST OF PARTIAL FULL EQUIVALENCE]	**0**	**-0.000933919**	**-0.003230958**

CFI= Comparative Fit Index; Γ_1= Gamma Hat fit index; Mc= McDonald Non-centrality index

Partial full equivalence may be concluded if the partial complete invariance multi-group measurement model in which the three non-invariant factor loadings (LA35, JS27 and RA6) are allowed to vary, the 31 non-invariant intercepts are allowed to vary, the four non-invariant error

variances (WH43, LA35, JS27, and REL9) are allowed to vary and the non-invariant variances and covariances are allowed to vary, does not fit practically significantly poorer than the configural invariance multi-group measurement model. As indicated in Table 5.76, for the partial complete invariance model in which all the latent variable variances were not constrained to be equal across gender groups and 15 latent variable pair covariances (refer to Table 5.70) were not constrained to be equal across gender groups, the change in CFI, Γ_1 and Mc was less than the critical thresholds of -0.01, -0.001 and -0.02 respectively[37]. This provided support for partial full equivalence.

5.21.1 DECISION ON THE SUCCESS OF PARTIAL FULL EQUIVALENCE

Partial full equivalence would adequately be supported when a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, and in which all parameters were constrained to be equal across the gender samples, but for the 3 slopes, 31 intercepts, 4 error variances and all variances and 15 covariances, which were allowed to vary, did not fit practically significantly poorer than a multi-group measurement model in which the structure of the model was constrained to be the same across gender groups, but all parameters were estimated freely. When the partial complete invariance model, which allowed for the non-invariant slopes of LA35, JS27 and RA6, the 31 non-invariant intercepts, the four non-invariant error variances of WH43, LA35, JS27 and REL9, and all 9 non-invariant variances and 15 non-invariant covariances to vary across gender groups, was compared to the configural invariance model, partial full equivalence was displayed as all three fit indices were below the critical thresholds.

Therefore, along with 56 of the 59 SWSI slopes being invariant, 29 of the 59 SWSI intercepts being invariant, 55 of the 59 error variances of the indicator variables being invariant, 21 covariances were found to be invariant, displaying adequate support for partial full equivalence.

[37] In the procedure used in this study, the equality constraints imposed on latent variable variances and latent variable pair covariances were lifted in consecutive partial complete invariance models by each time lifting an equality constraint on the most dissimilar variances and the most dissimilar covariances that were at that point still constrained. Consideration should, however, be given to the possibility of rank ordering latent variable variances differences and covariance differences together and lifting equality constraints on latent variable variances or covariances based on the rank ordered list.

CHAPTER 6
DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

6.1 INTRODUCTION

Work stress is a major, world-wide challenge to employee and organisational health. Occupational stress assessment is proposed as a method that could assist individuals in identifying their general level of stress and the possible sources of stress at work. If individuals and organisations are able to identify the possible sources of stress, pinpointing the problem area in the workplace can lead to planning and implementing interventions to improve employee wellbeing/psychological health, and ultimately performance.

With the backdrop of stronger demands placed on the gender appropriateness of psychological tests as well as in the interest of good workmanship, organisations and individuals should take note that their decisions could adversely affect the individuals being assessed, as well as the organisations employing the individuals, if the psychometric integrity of the assessments on which the decisions are based is in question. The decisions that are made on the basis of work stress information will have a substantial impact on both individuals and organisations. Therefore, it is essential to ensure that valid and reliable measures are used to measure employee wellness and more specifically occupational stress.

Any instrument used to assess stress attaches a specific connotative meaning (Kerlinger & Lee, 2000) to the stress construct. If the stress construct is defined as a multidimensional construct, specific latent stress dimensions would be distinguished in terms of this conceptualisation. Specific items are designed to serve as effect indicators (Hair et al., 2006) of these latent stress dimensions. This design intention is reflected in the scoring key of the instrument. The constitutive definition of the stress construct in conjunction with the design intention underlying the instrument implies a very specific measurement model.

Given the demands for gender appropriate psychological tests, a critical question is whether the measurement model reflecting the design intentions of the developers of the instrument fits the data of each gender group independently at least reasonably well. A further critical question is whether the measurement model parameters are the same across gender groups. The measurement models underlying stress instruments should be invariant and equivalent across gender, which would mean that the instrument measures the proposed areas of assessment (stress) in the same manner across the two gender groups. Should these measurement models not display transference

across the different groups (in this case gender), then, depending on the nature of the measurement model discrepancies, the test is ultimately testing different latent variables across the respective groups, or the test is measuring the same latent variable differently across the respective groups.

While invariance in terms of the number of factors and the associated pattern of factor loadings (i.e., configural invariance; Vandenberg & Lance, 2000) might satisfy one level of invariance, it certainly is not a sufficient condition for ensuring that the latent stress dimension in one group has been equivalently measured in the other group(s). The magnitude of the measurement model parameters could still differ across the different groups, and this could still imply non-invariance and non-equivalence in measurement. To be able to confidently interpret observed score differences between genders as indicative of latent score differences full measurement invariance needs to be indicated.

For invariance and equivalence to be observed, the identification and control of bias would be a necessary requirement. Theron (2011) describes bias as all systematic factors that could account for variance in observed test scores that cannot be accounted for in terms of the latent variable of interest. Construct, method, and item bias indicate where bias may originate. Dunbar et al. (2011) have provided a hierarchy that describe different levels of invariance and equivalence. Only once invariance and equivalence have been met at the highest level can observed scores be compared with confidence. Such confidence rests in the fact that the differences observed in scores between different groups are reflective of a true difference on the underlying latent variable, and not due to systematic group effects in measurement.

This then raises the question: how should measurement invariance and equivalence be evaluated? This study made use of a procedure that allows for specific hypotheses to be tested regarding measurement invariance and equivalence. Vandenberg and Lance (2000) raised the issue that measurement invariance research in organisational settings should be conducted routinely. Through confirmatory factor analytic procedures, the researcher was able to fit the measurement model implied by the constitutive definition of the stress construct (embodied in the SWSI) and the design intentions of the test publishers to data.

If reasonable measurement model fit along with significant ($p < .05$) and reasonably high completely standardised factor loadings would be found within each gender group, it would permit the within gender group use of the instrument to measure the stress construct as constitutively defined. Cross-gender group comparisons would, however, thereby not be sanctioned. An additional vital question is therefore whether the measurement model parameters are the same across the gender groups. In

examining these differences the confirmatory technique allows for placing increasing constraints on the model to determine at which level of constraint measurement invariance is being threatened (Vandenberg & Lance, 2000). This, however, does not answer the equivalence question of whether a multi-group model with fewer constraints imposed might not provide a more permissible explanation of the observed covariance matrices. Dunbar et al. (2011) indicate that invariance and equivalence should be addressed at five important corresponding levels: configural invariance, weak invariance and metric equivalence, strong invariance and scalar equivalence, strict invariance and conditional probability equivalence, and complete invariance and full equivalence. These coincide with the taxonomy of Van de Vijver and Leung (1997) and Vandenberg and Lance (2000), and the confirmatory technique allows for testing the measurement models as per the design intentions of the questionnaire authors/publishers.

In the taxonomy of measurement invariance and equivalence proposed by Dunbar et al. (2011), a finding of lack of invariance and/or lack of equivalence will result in the termination of testing as subsequent tests assume invariance on the previous level. This, however, seems an unnecessarily strict interpretation of invariance and equivalence both when viewing invariance and equivalence analysis from the perspective of measurement bias and cross-group comparative research. A follow-up procedure should be implemented, investigating partial invariance and partial equivalence. In testing for measurement invariance and equivalence, if the list of items, collectively, does not display invariance and/or equivalence across groups, the question arises as to which items are non-invariant (Cheung & Rensvold, 1999)? The appropriateness of partial invariance and partial equivalence is determined by whether non-invariant items can be accurately identified, and the extent of their departure from invariance (Cheung & Rensvold, 1999). This leads to the critical question of how should non-invariant constructs and/or items be identified?

Byrne et al.'s (1989) partial measurement invariance and equivalence procedure applies to factors that are configural invariant and therefore the non-invariant problem first emerges when weak invariance is imposed on the model. If the criterion for metric equivalence is not met (i.e. the weak invariance model fits significantly poorer than the less constrained configural invariance model), then additional tests are required to determine the sources of non-invariance (Cheung & Rensvold, 1999).

The procedure followed in this research involved (a) testing for partial metric equivalence per sub-scale, (b) examining the factor loadings of the configural invariance model for those sub-scales where a lack of partial metric equivalence was found to identify the greatest difference between gender groups (i.e., the most dissimilar factor loadings), and (c) testing for partial metric equivalence

when lifting the factor loading equality constraint one item at a time, starting with the item with the most dissimilar loadings across the two gender samples .

For a measurement model that assesses a multidimensional construct, each of the latent dimensions of the construct in the measurement model needs to be examined for invariance (Byrne et al., 1989; Cheung & Rensvold, 1999). The items for which the regression of X_i on ξ_j differs across gender samples can be limited to a specific sub-scale or can be scattered across two or more sub-scales. To narrow the search for the non-invariant items down to specific sub-scales each of the stress sub-scales in the model were examined for invariance (Cheung & Rensvold, 1999). A separate multi-group model was estimated in which the factor loadings associated with a specific latent dimension of the construct were constrained to be equal across gender groups, while the loadings associated with the other latent stress dimensions were freely estimated. The sub-scale in which statistical and/or practical significance was displayed indicated that at least one of the items within that subscale was non-invariant (Cheung & Rensvold, 1999). All non-invariant sub-scales were noted, and their items examined for invariance.

Following the identification of the non-invariant constructs, a series of tests was performed where a multi-group model was estimated in which the item with the most dissimilar factor loadings between gender groups was allowed to vary in order to identify non-invariant items. Once again, each of the constrained models was compared to the less constrained configural invariance model, via the calculation of the statistical and practical significance. If the change in the indices was significant, then that item was considered to be non-invariant. This procedure continued at each invariance and equivalence level. This procedure was used in correspondence with Dunbar et al.'s (2011) invariance and equivalence levels: partial weak invariance and partial metric equivalence, partial strong invariance and partial scalar equivalence, partial strict invariance and partial conditional probability equivalence, and partial complete invariance and partial full equivalence.

For this study, the Sources of Work Stress Inventory (SWSI; De Bruin & Taylor, 2005a) was chosen for the purpose of investigating the measurement invariance and equivalence issue. This questionnaire was chosen as the authors have indicated a preference for such a study to be conducted on this instrument, and in order to contribute to further research on this instrument.

Evidence on the psychometric integrity of the instrument is reported in the test manual (De Bruin & Taylor, 2005a). The validity and reliability analysis results reported in the manual was based on a South African sample. The literature reflects minimal South African studies that evaluated the reliability and construct validity of the SWSI. Moreover, none of the studies on the psychometric

integrity of the SWSI evaluated the fit, through confirmatory factor analysis, of the measurement model implied by the design intentions of the developers.

Establishing measurement invariance and equivalence of psychological measurement instruments is important and very relevant in the South African context. It is essential to establish whether the measurement tools used in South Africa do not display group-related measurement bias, with the ultimate goal to minimise systematic error in measurement, as far as possible. Therefore, it was decided that research should be conducted on the SWSI in a South African setting and the gender invariance and equivalence issue should be investigated. The test authors and local questionnaire distributors welcomed and authorised the research on this instrument. However, the researcher was not in a position to alter the instrument and the underlying measurement model in any way. This was due to the fact that the intellectual property rights for the instrument do not reside with the researcher, and that the researcher was not mandated by the test publisher to modify the design of the instrument in any way.

6.2 DISCUSSION

The objective of this study was to establish the extent to which the Sources of Work Stress Inventory measurement model may be considered measurement invariant and measurement equivalent across gender groups in South Africa. A series of measurement invariance and measurement equivalence tests were used to test the stability of the model parameter estimates in order to determine the source of the variance and the extent the measurement model may be considered measurement invariant and equivalent, partially invariant and equivalent, or not at all (Vandenberg & Lance, 2000).

Prior to fitting the measurement model, preliminary analyses that included item and dimensionality analyses as well as discriminant validity analyses were conducted on the respective sub-samples. Results for the item analyses revealed some problematic items that returned low inter-item correlations. Some items were problematic for one sample but not the other and vice versa. However, internal consistency figures for the SWSI sub-scales were above the benchmark of .80 set in this study for both samples. Nonetheless, reliability of the instrument for the South Africa sample has been established in this study.

Dimensionality analysis was performed via exploratory factor analysis on each SWSI sub-scale. The uni-dimensionality of the sub-scales was investigated as possible indicators of poor model fit in the subsequent CFA. An initial finding was that a limited number of sub-scales, in both the male and female sample, did not meet the uni-dimensionality assumption. In all three cases where uni-

dimensionality was not found (the Role Ambiguity and Work/Home Interface sub-scales in the male sample and the Role Ambiguity sub-scale in the female sample), a maximum of two factors were extracted. In all cases factor fission presented itself as a plausible explanation for the extraction of more than one factor. This would suggest that the Role Ambiguity sub-scale and the Work/Home Interface sub-scale contain meaningful sub-facets of sources of stress. However, when forcing the extraction of a single factor, the vast majority of items comprising the "split" sub-scales returned acceptable factor loadings (> .50), but also a large percentage of large residuals, which indicates the failure of the forced one factor solution to provide a credible explanation for the observed inter-item correlation matrix. Sub-scales that met the uni-dimensionality assumption, but that returned a too large percentage of large residuals (thereby failing to provide a credible explanation for the observed inter-item correlation matrix), were investigated for the possibility of a second factor. When the extraction of a second factor was forced for the Workload sub-scale in the male sample and for three of the sub-scales (Job Security sub-scale, Work/Home Interface sub-scale, and Workload sub-scale) in the female sample, a meaningful and credible factor structure emerged, suggesting that these sub-scales could be better explained by further sub-facets of the respective occupational stress scales.

Upon investigating the measurement model fit of the SWSI, the results provided support for the hypotheses that the measurement model fitted the data of both gender samples independently. Following satisfactory model fit for both gender samples further measurement invariance and equivalence analysis was justified. However, first discriminant validity was tested.

Discriminant validity of the SWSI measure was evaluated for each gender sample. Taking into account the two discriminant validity tests, the 95% confidence interval estimate and the average variance extracted versus the squared correlation, for the male sample, the results for the GEN, RA, CA, LA, and WH scales were problematic, as more variance in the indicators comprising the sub-scale was explained by measurement error than by the constructs the sub-scale items were designed to reflect. Furthermore, the RA and LA, REL and LA, LA and JS and WH and WL sub-scales were problematic in that the latent dimensions they measure shared more variance than the variance that at least one of the sub-scales explained of the construct they were tasked to reflect. Although, given the fact that none of the confidence intervals included unity, the latent dimensions measured by the SWSI may be considered qualitatively distinct, not all sub-scales sufficiently successfully tap into the unique aspects of the latent sources of stress they are meant to reflect. Based on these results it could be concluded that all the latent dimensions measured by the SWSI sub-scales can be described as qualitatively distinct, even though particular stress sub-scales do not successfully capture the

distinction that exists in the latent dimensions. Nonetheless, sufficient evidence was provided for discriminant validity in the male sample. Taking the tests for the female sample into account, the results for the RA, LA, and WH sub-scale were problematic, as more variance in the indicators comprising the sub-scale was explained by measurement error than by the constructs the sub-scale items were designed to reflect. Furthermore, the LA and REL, LA and CA, and LA and JS sub-scales were problematic in that the latent dimensions they measure shared more variance, than the variance that at least one of the sub-scales explained of the construct they were tasked to reflect. Given the fact that none of the confidence intervals included unity, the latent dimensions measured by the SWSI may be considered qualitatively distinct. The fact that for some of them ϕ^2_{ij} exceeded the average variance extracted by at least one of the sub-scales, meant that although each latent dimension may be considered sufficiently unique, some sub-scales do not altogether successfully capture the distinction that exists in the latent dimensions. Similar to the results for the male sample, it may be concluded that all the latent dimensions measured by the SWSI sub-scales can be described as qualitatively distinct, even though particular stress sub-scales do not fully succeed in capturing the unique part of the latent dimensions they were designed to represent. Overall it was concluded that sufficient evidence was provided for discriminant validity in the female sample.

The series of measurement invariance and measurement equivalence tests followed. The measurement model fitted successfully under the configural invariance condition, which required the structure of the model to be constrained across gender groups. This indicated that the measurement model reflects the same underlying construct across the gender groups, and signified that the different gender groups used the same conceptual frame of reference when they responded to the items. Following acceptable model fit on both gender samples independently and support for configural invariance, the measurement model also fitted successfully under the weak invariance condition, which required the structure and the factor loadings of the items on the latent variables to be constrained to be equal across gender groups. This provided support for the position that the items operated in approximately the same way across gender samples in the way they reflect the underlying latent variables they were meant to reflect (Dunbar et al., 2011). The finding of weak invariance was a satisfying outcome as it indicated that the item content was being perceived and interpreted in a similar manner across gender groups.

Further investigation revealed that the weak invariance multi-group model fitted significantly poorer than the configural invariance multi-group model, displaying a lack of metric equivalence. A procedure to explore which slopes were non-invariant was conducted next, resulting in the

identification of the slopes of three items (LA35, JS27, and RA6) displaying measurement non-invariance, supporting a finding of partial metric equivalence.

Taking the non-invariant slopes of the three items into account, the measurement model fitted successfully under the strong invariance condition, which indicated that the stance that the structure, the 56 regression slopes and all the intercepts of the items on the latent variables were the same across gender groups was permissible. Support was thus provided for the position that the items operate in approximately the same way across the gender samples in the way they reflect the underlying latent variables they are meant to reflect.

The test for scalar equivalence, however, did not obtain adequate support. The strong invariance multi-group model fitted significantly poorer than the configural invariance multi-group model. Further investigation allowed for the identification of the intercepts of the 31 items displaying measurement non-invariance, thereby finding support for partial scalar equivalence.

Taking the three non-invariant slopes and the 31 non-invariant intercepts into account, the measurement model fitted successfully under the strict invariance condition, which indicated that the position that the 56 regression slopes, the 28 intercepts and all the error variances of the indicator variables were the same across gender samples was plausible. Support was thus provided for the position that the respondents from the different gender groups respond to the SWSI in such a manner that no significant variance exists across samples in terms of error terms associated with the indicator variables.

However, further investigation revealed that the strict invariance multi-group model (which takes into account non-invariant items) did fit significantly poorer than the configural invariance multi-group model, displaying a lack of conditional probability equivalence. Subsequent investigation allowed for the identification of the error variances of the four items (WH43, LA35, JS27, and REL9) displaying measurement non-invariance, thereby finding support for partial conditional probability equivalence.

Taking the three non-invariant slopes, 31 non-invariant intercepts and the four non-invariant error variances into account, the measurement model fitted successfully under complete invariance. This indicated that the position that the 56 regression slopes, the 28 intercepts, the 55 error variances of the indicator variables, and all the latent variable variances and covariances was the same across gender samples was tenable. Support was thus provided for the position that the gender samples used equivalent ranges of the construct continuum to respond to the indicators reflecting the construct (Vandenberg & Lance, 2000).

However, the test for full equivalence did not obtain adequate support. The complete invariance multi-group model did fit significantly poorer than the configural invariance multi-group model. Further investigation allowed for the identification of the variances of all the subscales and the covariances of 15 pairs of items displaying measurement non-invariance, thereby finding support for partial full equivalence.

When reviewing the results of the CFA, only item RA6 displayed non-invariance in terms of both slope and intercept. When reviewing the preliminary analyses, this item (item RA6) did not appear to be problematic, but it had low factor loadings (< .40) on the two factors that were extracted in both the male and female sample. When a single factor was forced, item RA6 loaded satisfactorily on the single factor in both the male and female sample (> .50). When reviewing the results of the item analyses, EFA and CFA, none of the items in which the slope and/or intercept differences were evident were problematic items highlighted in Chapter 5. However, the non-invariant items that have been identified could be revisited in terms of content.

6.3 LIMITATIONS OF THE STUDY

It is important to acknowledge certain limitations of this study. Firstly, there was a lack of descriptive demographic information regarding the composition of the sample. Some of the observations made during the analysis could have been a function of the composition of the sample, thereby supporting the creation of further hypotheses to be tested. Also, in the event of obtaining further information regarding the composition of the sample, for example educational background, race or stage of employment, further invariance tests could be conducted. It would be necessary to conduct such further tests in order to obtain evidence that the SWSI instrument does not display other group-related measurement bias when taking other demographic information into account (such as race or stage of employment) other than gender, which was focused on in this research study.

Secondly, the results for invariance and equivalence across the gender samples presented in this study may not be interpreted to signify invariance and equivalence across other different groups within the target population, or across samples from other populations.

Thirdly, the procedure (Cheung & Rensvold, 1999) used to identify non-invariant subscales and items could potentially be a limitation. A number of methods exist and were discussed (refer to section 3.4), however the size of the questionnaire (i.e. number of items and latent variables) limited the options as the execution of such procedures would become quite complex and cumbersome.

6.4 RECOMMENDATIONS FOR RESEARCHERS AND PRACTITIONERS

Due to the limitation that invariance and equivalence across the gender samples used in this study may not be assumed to signify invariance and equivalence across other different groups within the target population or across samples from other populations, it is recommended that this study should be replicated across other samples from the target population in order to further establish the measurement invariance and measurement equivalence of the SWSI.

If possible, further measurement invariance and measurement equivalence tests should also be conducted on the SWSI across cultures, age, stage of employment etc. This is an important and relevant issue to address in South Africa. It is essential to establish whether the SWSI does not display group-related measurement bias in order to avoid making widespread generalisations and untested assumptions that will eventually do a disservice to the field of Industrial Psychology. Given the multicultural nature of the South African society, investigating the measurement invariance and measurement equivalence of the SWSI across cultural groups in South Africa would be vital and necessary. Without such evidence that the SWSI does not display group-related measurement bias, the use of the occupational stress assessment, and the decisions based on them, would seriously jeopardize the objectives the assessment intends on achieving. Further tests, however, would require the gathering of more data as well as more complete demographic information regarding the composition of the sample.

In the procedure used in this study, the equality constraints imposed on latent variable variances and latent variable pair covariances were lifted in consecutive partial complete invariance models by each time lifting an equality constraint on the most dissimilar variances and the most dissimilar covariances that were at that point still constrained. Consideration should, however be given to the possibility of rank ordering latent variable variances differences and covariance differences together and lifting equality constraints on latent variable variances or covariances based on the rank ordered list.

Lack of invariance was obtained on specific measurement model parameters that describe the relationships that exist between the items of the SWSI and the latent dimensions that the instrument measure. Decisions are, however, not based on the items of the SWSI. Decisions are based on the observed subscale scores of the SWSI, calculated from the items of each subscale. Difference in measurement model parameters across gender groups in an item in a subscale results in differences in the observed item score that cannot be explained by differences in the latent dimension being measured. These differences in the item score are purely brought about by differences in the regression relationship between the item and the latent trait. A subscale might

only contain a single invariant item. The critical question from the perspective of decision-making then is whether the biasing effect of this single item is maintained when the item responses are combined to form a composite dimension score or whether its effect is sufficiently diluted to make the dimension score effectively unbiased. More than one item in a subscale can, however, display differences in measurement model parameters. When the item responses are now combined to form a composite dimension score, the critical question from the perspective of decision-making is whether these differences in item parameters reinforce each other to create measurement bias in the dimension scores or whether the differences in item parameters oppose and cancel each other out to prevent measurement bias in the dimension scores. SEM does not seem to offer a procedure to investigate these questions. A measurement model in which each dimension is operationalised by a single dimension score would not be identified. Item response theory also does not offer a possibility of investigating this issue.

Lastly, an investigation into the structural invariance and equivalence of the SWSI is recommended. The constitutive meaning of a construct lies in the internal structure of the construct and the manner in which the construct and its dimensions are embedded in a nomological network of other constructs. The developers of SWSI designed the instrument so that specific items of the instrument reflect specific latent dimensions of the construct. The analyses reported in this study focused on the internal structure of the construct. The fact that the single-group gender-specific measurement models fitted the data is insufficient evidence to confidently conclude that the SWSI measures the stress construct as it was constitutively defined. By token of the fact the configural invariance model fitted the data provides insufficient evidence to confidently conclude that the SWSI measures the stress construct as it was constitutively defined across the two gender groups. Differences could still exist in the manner in which the stress dimensions measured by the SWSI are structurally embedded in the nomological network implied by the constitutive definition of stress.

Partial full equivalence was obtained for the SWSI, indicating that valid cross-group comparisons across gender groups can be conducted but, strictly speaking, only if the lack of invariance in specific measurement model parameters is taken into account. That raises the question how the practitioner can take the lack of invariance in specific measurement model parameters into account when making decisions based on dimension scores. Excluding biased items from the calculation of dimension scores is one option. It is, however, a wasteful option. Biased items still reflect information on the latent dimension of interest. They only do so in a manner that differs across gender groups. A second possibility that should be investigated is to calculate latent dimension

scores from the measurement model parameters that take into account differences that exist in measurement model parameters across gender groups.

LISREL offers the possibility of calculating latent scores from observed scores (Jöreskog, 2000). Jöreskog (2000, p. 4) derived the following matrix equation to obtain estimates of scores achieved on a latent variable (expressed as deviation scores):

$$\boldsymbol{\xi}^*_a = \boldsymbol{\kappa}^* + \mathbf{U}\mathbf{D}^{½}\mathbf{Z}^{-1}\mathbf{D}^{½}\mathbf{U}'\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}(\mathbf{x}^*_a - \boldsymbol{\tau} - \boldsymbol{\Lambda}\boldsymbol{\kappa}^*) \quad (4)$$

Symbols are defined in Jöreskog (2000). In the case of a single-group measurement model LISREL can be requested to utilise Equation 4 to estimate latent scores for the observations comprising the data set used to derive the model parameters. LISREL does not offer the possibility of utilising equation 4 to derive latent score estimates for subsequent samples. Neither does LISREL offer the possibility of using an extension of Equation 4 to derive latent score estimates for multi-group measurement models.

One possibility to overcome the difficulty preventing the calculation of latent scores for new samples is to output the necessary matrices from the fitted measurement model and to write dedicated software to calculate $\boldsymbol{\xi}^*_a$ from these matrices via Equation 4. An alternative solution is to fit the single-group measurement model to the new samples with all measurement model parameters fixed to the values obtained in the original analysis and to request the calculation of latent scores via LISREL.

The same procedure could possibly be extended to the multi-group measurement model by fitting separate gender-specific measurement models in which all measurement model parameters are fixed to the values obtained in the partial full equivalence multi-group model, and to request the calculation of latent scores via LISREL. A question is whether this procedure will work for data obtained for the small gender-diverse samples of respondents tested on each application of the SWSI. No parameters are estimated, and therefore it could be argued that the usual big sample requirement imposed on SEM falls away. An alternative avenue to explore is the possibility of developing latent score estimate norm tables from a simulated data set. The simulated data set will, however, have to make provision for all possible combinations of item score patterns across all subscales in both gender groups, and therefore will be an extremely large data set. This will also bring to the fore the question how to locate the corresponding data set in the norm table given a specific SWSI test protocol of a specific individual.

Although no parameters will be estimated in the fitted models, the fit of the model will nonetheless be evaluated via the usual spectrum of fit statistics. These would serve the useful function of commenting on the extent to which the multi-group measurement model derived in the validation study successfully transfers (or cross-validates) to the application sample.

6.5 CONCLUSION

The SWSI's measurement model CFA results presented in this study suggest that a fair amount of confidence can be placed in how the model could be replicated in the South African population. The measurement model fit was good, placing conviction in the interpretation and communication of results to questionnaire respondents. Furthermore, the SWSI indicated partial full equivalence, indicating that valid cross-group comparisons over gender groups can be conducted but, strictly speaking, only if the lack of invariance in specific measurement model parameters is taken into account.

Although the results also indicated that some differences existed when the measurement model was fitted to both samples simultaneously, and should be taken into account when examining the structural invariance of the SWSI, these results do not appear to threaten the conclusion that the SWSI is a credible measure of the stress construct it was intended to measure.

In conclusion, and despite the shortcomings outlined above, this measurement invariance and measurement equivalence study provides plausible evidence that the Sources of Work Stress Inventory measurement model demonstrates partial measurement invariance and equivalence across the gender groups in South Africa.

REFERENCES

Babbie, E., & Mouton, J. (2001). *The practice of social research*. Cape Town: Oxford University Press.

Beehr, T.A. (1995). *Psychological stress in the workplace.* London: Routledge.

Berry, J.W., Poortinga, Y.H., Segall, M.H., & Dasan, P.R. (2002). *Cross-cultural Psychology: Research and applications (2nd ed.).* Cambridge: Cambridge University Press.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch Model: Fundamental measurement in the human sciences.* Mahwah, NJ: Earlbaum.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A.Bollen & J.S. Long (Eds.), Testing structural equation models. Newbury Park: Sage Publications.

Byrne, B. (2001). *Structural equation modelling with AMOS: Basic concepts, application, and programming.* Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B.M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York: Springer Verlag.

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor co-variance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456-466.

Byrne, B.M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-cultural Psychology*, 34(2), 155-175.

Cheung, G.W., & Rensvold, R.B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management,* 25(1), 1-27.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit Indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233- 255.

Cooper, C.L., Dewe, P.J., & O'Driscoll, M.P. (2001). *Organisational Stress: a review and critique of theory, research, and applications.* United States of America: Sage Publications, Inc.

Cotton, P., & Hart, P.M. (2003). Occupational well-being and performance: A review of organisational health research. *Australian Psychologist,* 38 (2), 118-127.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace.

Davidson, J. (2000). *Econometric Theory* [Electronic version]. Blackwell Publishers.

De Beer, M. (2004). Use of Differential Item Functioning (DIF) analysis for Bias Analysis in Test Construction. *South African Journal of Industrial Psychology,* 30 (4), 52-58.

De Bruin, G.P. (2006). The dimensionality of the general work stress scale: a hierarchical exploratory factor analysis. *South African Journal of Industrial Psychology,* 32 (4), 68-75.

De Bruin, G.P., & Taylor, N. (2005). Development of the sources of work stress inventory. *South African Journal of Psychology,* 35(4), 748-765.

De Bruin, G.P., & Taylor, N. (2006a). *Sources of Work Stress Inventory: Technical Manual.* Johannesburg: Jopie van Rooyen & Partners.

De Bruin, G.P., & Taylor, N. (2006b). The job demand-control model of job strain across gender. *South African Journal of Industrial Psychology*, 32 (1), 66-73.

Diamantopoulos, A., & Siguaw, J.A. (2000). *Introducing LISREL*. London: SAGE Publications, Inc.

Donnelly, C. (2009). A Multi-Group Structural Equation Modelling Investigation of the Measurement Invariance of the Cambell Interest and Skill Survey (CISS) Across Gender Groups in South Africa. Unpublished Masters thesis, University of Stellenbosch, Cape Town.

Doyle, C., & Hind, P. (1998). Occupational stress, burnout and job status in female academics. *Gender, Work & Organisation*, 5, 67-82.

Dunbar, H., Theron, C., & Spangenberg, H. (2011). A cross-validation study of the Performance Index. *Management Dynamics*, 20(3), 2-24.

Dunbar-Isaacson, H. (2006). An investigation into the measurement invariance of the performance index. Stellenbosch: Unpublished dissertation, University of Stellenbosch.

Du Toit, M., & Du Toit, S. (2001). Interactive LISREL: User's guide. Lincolnwood, IL: Scientific Software International.

Edwards, A.L. (1957). *Techniques of attitude scale construction.* New York: Appleton- Century-Crofts.

Edwards, A.L. (1970). *The measurement of personality traits by scales and inventories.* New York: Holt, Rinehart & Winston.

Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research*, 63, 324-327.

Ferrando, P.J., & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicológica*, 21, 301-323.

Foxcroft, C., & Roodt, G. (2005*). An introduction to psychological assessment in the South African context (2nd ed.)*. Cape Town: Oxford University Press.

Görgens-Ekermans, G., & Brand, T. (2012). Emotional Intelligence as a moderator in the stress-burnout relationship: a questionnaire study on nurses. *Journal of Clinical Nursing*, 21, 2275-2285.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., & Tatham, R.L. (2006). *Multivariate data analysis (6th ed.)*. Upper Saddle River, NJ: Pearson Education Inc.

Hamidi, Y., & Eivazi, Z. (2010). The relationship among employees' job stress, job satisfaction and the organisational performance of Hamadan urban health centres. *Social Behaviour and Personality*, 2010, 38 (7), 963-968.

Hoffman. D.A., & Tetrick, L.E. (2003). The etiology of the concept of health: Implications for "organising" individual and organisational health. In D.A. Hofmann & L.E. Tetrick (Eds), *Health and Safety in Organisations: a multilevel perspective,* (pp. 1 -21). San Francisco: Jossey Bass.

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research,* 18, 117-144.

Hu, L.T., & Bentler, P.M. (1995). Evaluating model fit. In R.H. Hoyle (Ed.), *Structural equation modelling: Concepts, issues and applications.* Thousand Oaks, California: Sage Publications.

Hulin, C.L., Drasgow F., & Parsons C.K. (1983). *Item response theory: Application to psychological measurement*. Homewood, Ill.: Jones-Irwin Publishers.

Johnson, J.V., & Hall, EM. (1988). Job strain, work place social support, and cardiovascular disease: a cross-sectional study of a random sample of the Swedish working population. *American Journal of Public Health*, 78, 1336-1342.

Johnson, T.J., Kulesa, P., Cho, Y.I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology,* 36 (2), 264-277.

Johnson, T.P., O'Rouke, D., Chavez, N., Sudman, S., Warnecke, R., & Lacey, L. (1997). Social Cognition and responses to survey questions among culturally diverse populations. In L. Lyberg, P. Biemer, M. Collings, E. de Leeuuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), Survey measurement and process quality (pp. 87-113). New York: John Wiley & Sons.

Jöreskog, K.G. (2000). *Latent scores and their use*. United States of America: Scientific Software International, Inc.

Jöreskog, K.G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. United States of America: Scientific Software International, Inc.

Jöreskog, K.G., & Sörbom, D. (1996a). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (1996b*). PRELIS 2: User's reference guide*. Chicago: Scientific Software International.

Karasek, R.A., Jr. (1979). Job demands, job decision latitude and mental strain: Implications for job redesign. *Administrative Science Quarterly*, 24, 285-308.

Kelloway, E.K. (1998). *Using LISREL for structural equation modelling: A researcher's guide*. Thousand Oaks, CA: SAGE Publications, Inc.

Kerlinger, F.N., & Lee, H.B. (2000). *Foundations of behavioural research (4th ed.)*. Forth Worth, TX: Harcourt College Publishers.

Kim, S., & Hagtvet, K.A. (2003). The impact of misspecified item parcelling on representing latent variables in covariance structure modelling: a simulation study. *Structural Equation Modeling*, 10 (1), 101-127.

Kline, R.B. (2005). *Principles and practice of structural equation modelling (2nd ed).* New York: The Guilford Press.

Leka, S., Griffiths, A., & Cox, T. (2003). Work, Organisations and Stress: Systematic Problem Approach for Employers, Managers and Trade Union Representatives. In Institute of Work, Health and Organisations, Protecting Workers Health Series No.3 [Electronic version]. Geneva: World Health Organisation.

Lubke, G.H., & Muthén, B.O. (2004). Applying multi-group confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling,* 11(4), 514-534.

MacCallum, R.C. (1995). Model specification: procedures, strategies and related issues. In Hoyle, R.H. (Ed.), Structural equation modelling: Concepts, issues and applications. Thousand Oaks, CA: SAGE Publications, Inc.

MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modelling. *Psychological Methods*, 1 (2), 130-149.

Marsh, H.W., Hau, K., Balla, J.R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioural Research, 33*(2), 181-220.

Marsh, H., & Hocevar, D. (1985). The application of confirmatory factor analysis to the study of self-concept: First and higher order factor structures and their invariance across age groups. *Psychological Bulletin, 97*, 562-82.

McDonald, R.P. (1999). *Test theory*. Mahwah, NJ: Erlbaum.

McLean, A.A. (1979). *Work stress*. Philippines: Addison-Wesley Publishing Company Inc.

Meade, A.W., & Kroustalis (2006). Problems with item parcelling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, 9 (3), 369- 403.

Meade, A.W., & Lautenschleager, G.J. (2004). A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence. *Organizational Research Methods*, 7, 361-388.

Mels, G. (2003). *A workshop on structural equation modelling with LISREL 8.54 for Windows*. Chicago, IL: Scientific Software International.

Mels, G. (2007). *Statistical modeling with LISREL 8.80 for Windows*. Chicago: Scientific Software International.

Mels, G. (2010). *A workshop on structural equation modelling with LISREL 9 for Windows*. Chicago, IL: Scientific Software International.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika,* 58, 525-543.

Murphy, K.R., & Davidshofer, C.O. (2005). *Psychological testing: principles and applications* (6th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Muthén, B., & Kaplan, D. 1985. A comparison of some methodologies for factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38: 171-189.

Nunnally, J.C. (1978*). Psychometric theory*. New York: McGraw-Hill.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory (3rd ed.).* New York: McGraw-Hill.

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modelling. *European Journal of Operational Research*, 151, 53-79.

Patterson, H., & Uys, K. (2005). Critical issues in psychological test use in the South African workplace. SA Journal of Industrial Psychology, 31(3), 12-22.

Preacher, K.J., & Coffman, D.L. (2006, May). Computing power and minimum sample size for RMSEA (Computer software). Available from http://quantpsy.org/.

Sass, D.A., & Smith, P.L. (2006). The effects of parcelling unidimensional scales on structural parameter estimates in structural equation modelling. *Structural Equation Modeling*, 13 (4), 566-586.

Schlebusch, L. (1998). Recent advances in stress research and implications for health and well-being. In L. Schlebusch (Ed), *South Africa beyond transition: Psychological well-being*. Proceedings of the Third Annual Congress of the Psychological Society of South Africa (pp. 265-283). Pretoria: Psychological Society of South Africa.

Spangenberg, H.H., & Theron C.C. (2005). Promoting ethical follower behaviour through leadership of ethics: The development and psychometric evaluation of the ethical leadership inventory (ELI). *SA Journal of Business Management*, 36 (2), 1-18.

Spector, P.E. (2003). Individual difference in health and well-being in organisations. In D. A. Hofmann & L. E. Tetrick (Eds), *Health and safety in organisations: a multilevel perspective*, (pp. 29 – 49). San Francisco: Jossey Bass.

Steenkamp, J.B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national research. *Journal of consumer Research*, 25, 28-90.

Stevens, S.S. (1946). *On the theory of scales of measurement. Science, 103 (2684), 677-680.*

Stewart, D. (2001). Factor analysis. *Journal of Consumer Psychology*, 10 (1&2), 75- 82.

Tabachnick, B.G., & Fidell, L.S. (1989). *Using multivariate statistics*. New York: Harper Collins Publishers.

Tabachnick, B.G., & Fidell, L.S. (2007*). Using multivariate statistics (5th ed.)*. Boston: Pearson Education Inc.

Theorell, T. & Karasek, R.A. (1996). Current issues relating to job strain and cardiovascular disease research. *Journal of Occupational Health Psychology*, 1, 9-26.

Theron, C.C. (2007). Confessions, scapegoats and flying pigs: psychometric testing and the law. *South African Journal of Industrial Psychology*, 33(1), 102-117.

Theron, C.C. (2011). Intermediate statistics and computer usage. Unpublished class notes (Industrial Psychology 815), University of Stellenbosch.

Vandenberg, R.J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods,* 5(2), 139-158.

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods,* 3, 4-69.

Van der Doef, M., & Maes, S. (1999). The job demand-control (-support) model and psychological well-being: A review of 20 years of empirical research. *Work & Stress*, 13, 87-114.

Van de Vijver, F.J.R., & Leung, K. (1997). *Methods and data analysis for cross cultural research.* Thousand Oaks, CA: Sage Publications Inc.

Van de Vijver, F.J.R., & Poortinga Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment,* 13, 21-29.

Van de Vijver, F.J.R., & Rothman, S. (2004). Assessment in Multicultural Groups: The South African Case. *South African Journal of Industrial Psychology,*30 (4), 1-7.

Van de Vijver, F.J.R., & Tanzer, N.K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Européenne de Psychologie Appliquée*, 54, 119-135.

Vermeulen, M., & Mustard, C. (2000). Gender differences in job strain, social support at work, and psychological distress. *Journal of Occupational Health Psychology*, 5, 428-440.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis*. Chicago,Ill: MESA Press.